Statistical Methods in the Atmospheric Sciences

DANIEL S. WILKS

Statistical Methods in
the Atmospheric Sciences

This is Volume 59 in
INTERNATIONAL GEOPHYSICS SERIES
A series of monographs and textbooks
Edited by RENATA DMOWSKA and JAMES R. HOLTON

A list of the books in this series appears at the end of this volume.

Statistical Methods in the Atmospheric Sciences

An Introduction

Daniel S. Wilks
DEPARTMENT OF SOIL, CROP, AND ATMOSPHERIC SCIENCES
CORNELL UNIVERSITY
ITHACA, NEW YORK

ACADEMIC PRESS
San Diego New York Boston London Sydney Tokyo Toronto

The cover art is a schematic illustration of concepts in ensemble forecasting, plotted in terms of a two-dimensional phase space (see Fig. 6.21 for more details).

This book is printed on acid-free paper. ♾

Copyright © 1995 by ACADEMIC PRESS, INC.

All Rights Reserved.
No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Academic Press, Inc.
A Division of Harcourt Brace & Company
525 B Street, Suite 1900, San Diego, California 92101-4495

United Kingdom Edition published by
Academic Press Limited
24-28 Oval Road, London NW1 7DX

Library of Congress Cataloging-in-Publication Data

Wilks, Daniel S.
Statistical methods in the atmospheric sciences : an introduction / by Daniel S. Wilks.
p. cm. -- (International geophysics series : v. 59)
Includes index.
ISBN 0-12-751965-3
1. Atmospheric physics--Statistical methods. I. Title. II. Series.
QC874.5.W55 1995
551.6'3'01195--dc20 94-31848
CIP

PRINTED IN THE UNITED STATES OF AMERICA
95 96 97 98 99 00 MM 9 8 7 6 5 4 3 2 1

Contents

Preface xi

Chapter 1 Introduction

1.1 What Is Statistics? 1
1.2 Descriptive and Inferential Statistics 2
1.3 Uncertainty about the Atmosphere 2
1.4 An Aside on Notation 4

Chapter 2 Review of Probability

2.1 Background 6
2.2 The Elements of Probability 7
2.2.1 Events 7
2.2.2 The Sample Space 7
2.2.3 The Axioms of Probability 8
2.3 The Meaning of Probability 9
2.3.1 Frequency Interpretation 9
2.3.2 Bayesian (Subjective) Interpretation 9
2.4 Some Properties of Probability 10
2.4.1 Domain, Subsets, Complements, and Unions 10
2.4.2 Conditional Probability 12
2.4.3 Independence 14
2.4.4 Law of Total Probability 16
2.4.5 Bayes' Theorem 17
Exercises 19

Chapter 3 Empirical Distributions and Exploratory Data Analysis

3.1 Background 21
3.1.1 Robustness and Resistance 21
3.1.2 Quantiles 22
3.2 Numerical Summary Measures 24
3.2.1 Location 24
3.2.2 Spread 25
3.2.3 Symmetry 26

3.3 Graphical Summary Techniques . . . 27
3.3.1 Stem-and-Leaf Display . . . 27
3.3.2 Boxplots . . . 29
3.3.3 Schematic Plots . . . 30
3.3.4 Other Boxplot Variants . . . 33
3.3.5 Histograms . . . 33
3.3.6 Cumulative Frequency Distributions . . . 34
3.4 Reexpression . . . 36
3.4.1 Power Transformations . . . 37
3.4.2 Standardized Anomalies . . . 41
3.5 Exploratory Techniques for Paired Data . . . 44
3.5.1 Scatterplots . . . 44
3.5.2 Pearson ("Ordinary") Correlation . . . 45
3.5.3 Rank Correlation . . . 50
3.5.4 Serial Correlation . . . 51
3.5.5 Autocorrelation Function . . . 53
3.6 Exploratory Techniques for Higher-Dimensional Data . . . 54
3.6.1 The Correlation Matrix . . . 55
3.6.2 The Scatterplot Matrix . . . 57
3.6.3 Correlation Maps . . . 59
Exercises . . . 62

Chapter 4 Theoretical Probability Distributions

4.1 Background . . . 64
4.1.1 What Is a Theoretical Distribution? . . . 65
4.1.2 Parameters versus Statistics . . . 65
4.1.3 Discrete versus Continuous Distributions . . . 66
4.2 Discrete Distributions . . . 66
4.2.1 Binomial Distribution . . . 66
4.2.2 Geometric Distribution . . . 70
4.2.3 Poisson Distribution . . . 71
4.3 Statistical Expectations . . . 74
4.3.1 Expected Value of a Random Variable . . . 74
4.3.2 Expected Value of a Function of a Random Variable . . . 74
4.4 Continuous Distributions . . . 76
4.4.1 Distribution Functions and Expected Values . . . 76
4.4.2 Gaussian Distribution . . . 79
4.4.3 Gamma Distribution . . . 86
4.4.4 Weibull Distribution . . . 93
4.4.5 Lognormal Distribution . . . 95
4.4.6 Beta Distribution . . . 95
4.4.7 Gumbel Distribution . . . 97
4.5 Multivariate Probability Distributions . . . 99
4.5.1 Bivariate Normal Distribution . . . 99
4.5.2 Multivariate Normal Distribution . . . 102
4.6 Qualitative Assessments of the Goodness of Fit . . . 104
4.6.1 Superposition of a Fitted Theoretical Distribution and Data Histogram . . . 104
4.6.2 Probability Plots . . . 106
4.7 Parameter Fitting Using Maximum Likelihood . . . 108
Exercises . . . 111

Chapter 5 Hypothesis Testing

5.1 Background . . . 114
5.1.1 Parametric versus Nonparametric Tests . . . 114
5.1.2 The Sampling Distribution . . . 115
5.1.3 The Elements of Any Hypothesis Test . . . 115
5.1.4 Test Levels and p Values . . . 116
5.1.5 Error Types and the Power of a Test . . . 116
5.1.6 One-Sided versus Two-Sided Tests . . . 117
5.1.7 Confidence Intervals: Inverting Hypothesis Tests . . . 118
5.2 Some Parametric Tests . . . 121
5.2.1 One-Sample t Test . . . 121
5.2.2 Tests for Differences of Mean under Independence . . . 122
5.2.3 Tests for Differences of Mean under Serial Dependence . . . 125
5.2.4 Goodness-of-Fit Tests . . . 129
5.2.5 Likelihood Ratio Test . . . 135
5.3 Nonparametric Tests . . . 137
5.3.1 "Classical" Nonparametric Tests for Location . . . 138
5.3.2 Resampling Tests . . . 145
5.4 Field Significance and Multiplicity . . . 151
5.4.1 The Multiplicity Problem for Independent Tests . . . 151
5.4.2 Field Significance Given Spatial Correlation . . . 152
Exercises . . . 157

Chapter 6 Statistical Weather Forecasting

6.1 Background . . . 159
6.2 Review of Least-Squares Regression . . . 160
6.2.1 Simple Linear Regression . . . 160
6.2.2 Distribution of the Residuals . . . 163
6.2.3 The Analysis-of-Variance Table . . . 165
6.2.4 Goodness-of-Fit Measures . . . 166
6.2.5 Sampling Distributions of the Regression Coefficients . . . 168
6.2.6 Examining Residuals . . . 171
6.2.7 Prediction Intervals . . . 175
6.3 Objective Forecasts—Without NWP . . . 181
6.3.1 Stratification and Compositing . . . 182
6.3.2 When the Predictand Is a Probability . . . 182
6.3.3 Predictor Selection . . . 185
6.3.4 Screening Regression . . . 188
6.3.5 Stopping Rules . . . 189
6.3.6 Cross-Validation . . . 194
6.3.7 Analog Forecasting . . . 198
6.4 Objective Forecasts—With NWP . . . 199
6.4.1 Perfect Prog Forecasts . . . 200
6.4.2 Model Output Statistics (MOS) Forecasts . . . 201
6.4.3 Comparisons between the Classical, Perfect Prog, and MOS Approaches . . . 202
6.4.4 Operational MOS Forecasts . . . 209
6.5 Probabilistic Field (Ensemble) Forecasts . . . 210
6.5.1 Stochastic Dynamic Forecasts . . . 211

6.5.2 Ensemble Forecasting . . . 213
6.5.3 Choosing Initial Ensemble Members . . . 215
6.5.4 Ensemble Averaging . . . 216
6.5.5 Dispersion of the Ensemble . . . 217
6.6 Subjective Probability Forecasts . . . 221
6.6.1 The Subjective Distribution . . . 223
6.6.2 Credible Interval Forecasts . . . 225
6.6.3 Assessing Discrete Probabilities . . . 227
6.6.4 Assessing Continuous Distributions . . . 229
Exercises . . . 230

Chapter 7 Forecast Verification

7.1 Background . . . 233
7.1.1 Purposes of Forecast Verification . . . 233
7.1.2 Factorization of the Joint Distribution of Forecasts and Observations . . . 234
7.1.3 Scalar Attributes of Forecast Performance . . . 235
7.1.4 Forecast Skill . . . 237
7.2 Categorical Forecasts of Discrete Predictands . . . 238
7.2.1 The Contingency Table . . . 238
7.2.2 Accuracy Measures for Binary Forecasts . . . 239
7.2.3 Bias . . . 241
7.2.4 Accuracy Measures for Multicategory Forecasts . . . 242
7.2.5 Conversion of Probabilistic to Categorical Forecasts . . . 244
7.2.6 Skill Measures for Contingency Tables . . . 248
7.3 Categorical Forecasts of Continuous Predictands . . . 250
7.3.1 Conditional Quantile Plots . . . 251
7.3.2 Scalar Accuracy Measures . . . 251
7.3.3 Skill Scores . . . 255
7.3.4 Skill in Probability Space . . . 257
7.4 Probability Forecasts . . . 258
7.4.1 Dichotomous Events . . . 259
7.4.2 Brier Score . . . 259
7.4.3 Algebraic Decomposition of the Brier Score . . . 260
7.4.4 The Attributes Diagram . . . 263
7.4.5 The Reliability Diagram . . . 265
7.4.6 Hedging and Strictly Proper Scoring Rules . . . 267
7.4.7 Probability Forecasts for Multicategory Events . . . 268
7.4.8 Ranked Probability Score . . . 269
7.5 Categorical Forecasts of Fields . . . 272
7.5.1 S1 Score . . . 274
7.5.2 Mean-Squared Error . . . 275
7.5.3 Anomaly Correlation . . . 277
Exercises . . . 281

Chapter 8 Time Series

8.1 Background . . . 284
8.1.1 Stationarity . . . 284
8.1.2 Time-Series Models . . . 285
8.1.3 Time-Domain versus Frequency-Domain Approaches . . . 286

8.2 Time Domain. I. Discrete Data . . . 287
8.2.1 Markov Chains . . . 287
8.2.2 Two-State, First-Order Markov Chains . . . 288
8.2.3 Test for Independence versus First-Order Serial Dependence . . . 292
8.2.4 Some Applications of Two-State Markov Chains . . . 295
8.2.5 Multistate Markov Chains . . . 297
8.2.6 Higher-Order Markov Chains . . . 298
8.2.7 Deciding among Alternative Orders of Markov Chains . . . 300
8.3 Time Domain. II. Continuous Data . . . 302
8.3.1 First-Order Autoregression . . . 302
8.3.2 Higher-Order Autoregression . . . 308
8.3.3 Order Selection Criteria . . . 313
8.3.4 The Variance of a Time Average . . . 315
8.3.5 Autoregressive Moving-Average Models . . . 318
8.3.6 Simulation and Forecasting with Time-Domain Models . . . 319
8.3.7 Multivariate Autoregressions . . . 323
8.4 Frequency Domain. I. Harmonic Analysis . . . 325
8.4.1 Cosine and Sine Functions . . . 325
8.4.2 Representing a Simple Time Series with a Harmonic Function . . . 326
8.4.3 Estimation of the Amplitude and Phase of a Single Harmonic . . . 330
8.4.4 Higher Harmonics . . . 333
8.4.5 Harmonic Dials . . . 336
8.4.6 The Harmonic Functions as Uncorrelated Regression Predictors . . . 339
8.5 Frequency Domain. II. Spectral Analysis . . . 341
8.5.1 The Periodogram or Fourier Line Spectrum . . . 341
8.5.2 Computing Spectra . . . 346
8.5.3 Aliasing . . . 347
8.5.4 Sampling Properties of Spectral Estimates . . . 350
8.5.5 Theoretical Spectra of Autoregressive Models . . . 352
Exercises . . . 357

Chapter 9 Methods for Multivariate Data

9.1 Background . . . 359
9.2 Matrix Algebra Notation . . . 360
9.2.1 Vectors . . . 360
9.2.2 Matrices . . . 362
9.3 Principal-Component (EOF) Analysis . . . 373
9.3.1 Basics of PCA . . . 374
9.3.2 Truncation of the Principal Components . . . 378
9.3.3 How Many Principal Components Should Be Retained? . . . 380
9.3.4 PCA Based on the Covariance Matrix versus the Correlation Matrix . . . 383
9.3.5 Application of PCA to Fields . . . 386
9.3.6 Rotation of the Eigenvectors . . . 393
9.3.7 The Varied Terminology of PCA . . . 394
9.3.8 Scaling Conventions in PCA . . . 397
9.4 Canonical Correlation Analysis . . . 398
9.4.1 Canonical Variates . . . 399
9.4.2 Computation of CCA . . . 400
9.4.3 CCA Applied to Fields . . . 403
9.4.4 Combining PCA and CCA . . . 403

9.5 Discriminant Analysis 408
9.5.1 Linear Discriminant Analysis 409
9.5.2 Multiple Discriminant Analysis 415
9.6 Cluster Analysis 419
9.6.1 Distance Measures and Clustering Methods 420
9.6.2 The Dendrogram, or Tree Diagram 423
9.6.3 How Many Clusters? 424
Exercises 427

Appendix A: Example Data Sets 429
Appendix B: Probability Tables 432
Appendix C: Answers to Exercises 439
References 444
Index 455
International Geophysics Series 465

Preface

This text is intended as an introduction to the application of statistical methods to atmospheric data. The structure of the book is based on a course that I teach at Cornell University. The course primarily serves upper-division undergraduates and beginning graduate students, and the level of the presentation here is targeted to that audience. It is an introduction in that many topics relevant to the use of statistical methods with atmospheric data are presented, but nearly all of them could have been treated at greater length and in more detail. The text will provide a working knowledge of some basic statistical tools sufficient to make accessible the more complete and advanced treatments available elsewhere.

It has been assumed that the reader has completed a first course in statistics, but basic statistical concepts are reviewed before being used. The book might be regarded as a second course in statistics for those interested in atmospheric or other geophysical data. For the most part, a mathematical background beyond first-year calculus is not required. A background in atmospheric science is also not necessary, but it will help the reader to appreciate the flavor of the presentation. Many of the approaches and methods are applicable to other geophysical disciplines as well.

In addition to serving as a text, this book will be a useful reference both for researchers and for more operationally oriented practitioners. Much has changed in atmospheric science since the 1958 publication of the classic *Some Applications of Statistics to Meteorology,* by Hans A. Panofsky and Glenn W. Brier, and no really suitable replacement has since appeared. For this audience, my explanations of statistical tools that are commonly used in atmospheric research will increase the accessibility of the literature and improve readers' understanding of what their data sets mean.

Finally, I acknowledge the help I received from Rick Katz, Allan Murphy, Art DeGaetano, Richard Cember, Martin Ehrendorfer, Tom Hamill, Matt Briggs, and Pao-Shin Chu. Their thoughtful comments on early drafts have added substantially to the clarity and completeness of the presentation.

Daniel S. Wilks

This Page Intentionally Left Blank

Chapter 1 Introduction

1.1 What Is Statistics?

This text is concerned with the use of statistical methods in the atmospheric sciences, specifically, in the various specialties within meteorology and climatology. Students (and others) often resist statistics, and the subject is perceived by many to be the epitome of dullness. Before the advent of cheap and widely available computers this negative view had some basis, at least with respect to applications of statistics involving the analysis of data. Performing hand calculations, even with the aid of a scientific pocket calculator, was, indeed, tedious, mind-numbing, and time-consuming. Although the capacity of ordinary personal computers on today's desktops is comparable to the fastest mainframe computers of 30 years ago, apparently many people have not noticed that the age of computational drudgery in statistics has long passed. In fact, some important and powerful statistical techniques were not even practical before the abundant availability of computing. Even when liberated from hand calculations, statistics is sometimes perceived as dull by people who do not appreciate its relevance to scientific problems. Hopefully, this text will help provide that appreciation, at least with respect to the atmospheric sciences.

Fundamentally, statistics is concerned with uncertainty. Evaluating and quantifying uncertainty, as well as making inferences and forecasts in the face of uncertainty, are all parts of statistics. It should not be surprising, then, that statistics has many roles to play in the atmospheric sciences, since it is the uncertainty in atmospheric behavior that makes the atmosphere interesting. For example, many people are fascinated with weather forecasting. Weather forecasting remains interesting precisely because of the uncertainty that is intrinsic to the problem. If it were possible to make perfect forecasts even one day into the future (i.e., if there were no uncertainty involved), the practice of meteorology would be very dull.

1.2 Descriptive and Inferential Statistics

It is convenient, although somewhat arbitrary, to divide statistics into two broad areas: descriptive statistics and inferential statistics. Both are relevant to the atmospheric sciences.

Descriptive statistics relates to the organization and summarization of data. The atmospheric sciences are awash with data. Worldwide, operational surface and upper-air observations are routinely taken at thousands of locations in support of weather forecasting activities. These are supplemented with aircraft, radar, profiler, and—increasingly—satellite data. Observations of the atmosphere specifically for research purposes are less widespread, but often involve very dense sampling in time and space. In addition, models of the atmosphere consisting of numerical integration of equations describing atmospheric dynamics produce yet more numerical output for both operational and research purposes.

As a consequence of these activities, one is often confronted with extremely large batches of numbers that, we hope, contain information about natural phenomena of interest. It can be a nontrivial task just to make some preliminary sense of such data sets. It is typically necessary to organize the raw data, and to choose and implement appropriate summary representations. When the individual data values are too numerous to be grasped individually, a summary that nevertheless portrays important aspects of their variations—a *statistical model*—can be invaluable in understanding the data. It may be worth emphasizing that it is not the purpose of descriptive data analyses to play with numbers. Rather, these analyses are undertaken because it is known, suspected, or hoped that the data contain information about a natural phenomenon of interest.

Inferential statistics is traditionally understood to consist of methods and procedures used to draw conclusions regarding underlying processes that generate the data. Thiébaux and Pedder (1987) express this point prosaically, when they state that statistics is "the art of persuading the world to yield information about itself." There is a kernel of truth here: Our physical understanding of atmospheric phenomena comes in part through statistical manipulation and analysis of data. In the context of the atmospheric sciences it is probably sensible to interpret inferential statistics a bit more broadly as well, and include statistical weather forecasting. By now this important field has a long tradition, and it is an integral part of operational weather forecasting at meteorological centers throughout the world.

1.3 Uncertainty about the Atmosphere

Underlying both descriptive and inferential statistics is the notion of uncertainty. If atmospheric processes were constant, or strictly periodic, describing them mathematically would be easy. Weather forecasting would also be easy, and me-

teorology would be boring. Of course, the atmosphere exhibits variations and fluctuations that are irregular. This uncertainty is the driving force behind the collection and analysis of the large data sets referred to in the previous section. It also implies that weather forecasts are inescapably uncertain. The weather forecaster predicting a particular temperature on the following day is not at all surprised (and perhaps is even pleased) if the subsequently observed temperature is different by a degree or two. In order to deal quantitatively with uncertainty it is necessary to employ the tools of probability, which is the mathematical language of uncertainty.

Before reviewing the basics of probability, it is worthwhile to examine why there is uncertainty about the atmosphere. After all, we now have large, sophisticated computer models that represent the physics of the atmosphere. Such models are used routinely for forecasting the future evolution of the atmosphere. These models are deterministic: They do not represent uncertainty. Once supplied with a particular initial atmospheric state (e.g., winds, temperatures, and humidities) and boundary forcings (e.g., solar radiation and, for some atmospheric models, such fields as sea–surface temperatures), each will produce a single particular result. Rerunning the model with the same inputs will not change that result.

In principle, these atmospheric models could provide forecasts with no uncertainty, but do not, for two reasons. First, even though the models can be very impressive and give quite good approximations to atmospheric behavior, they are not complete and true representations of the governing physics. An important and essentially unavoidable cause of this problem is that some relevant physical processes operate on scales too small to be represented by these models.

Even if all the relevant physics could be included in atmospheric models, however, we still could not escape the uncertainty because of what has come to be known as *dynamical chaos*. This is a problem "discovered" by an atmospheric scientist (Lorenz, 1963), and it has become the death knell for the dream of perfect (uncertainty-free) weather forecasts. Lorenz (1993) has also provided a very readable introduction to this subject. Simply and roughly put, the time evolution of a nonlinear, deterministic dynamical system (e.g., the equations of atmospheric motion, or of the atmosphere itself) is very sensitive to the initial conditions of the system. If two realizations of such a system are started from two only very slightly different initial conditions, the two solutions will eventually diverge markedly. For the case of weather forecasts, imagine that one of these systems is the real atmosphere and the other is a perfect mathematical model of the physics governing the atmosphere. Since the atmosphere is always incompletely observed, it will never be possible to start the mathematical model in exactly the same state as the real system. So even if the model is perfect, it will still be impossible to calculate what the atmosphere will do indefinitely far into the future. Therefore, *deterministic* forecasts of future atmospheric behavior will always be uncertain, and probabilistic methods will always be needed to adequately describe that behavior. Even if

the atmosphere is not fundamentally a random system, for many practical purposes it might as well be. This thought has been put very nicely by Zeng *et al.* (1993), who observe "Just as relativity eliminated the Newtonian illusion of absolute space and time, and as quantum theory eliminated the Newtonian and Einsteinian dream of a controllable measurement process, chaos eliminates the Laplacian fantasy of long-term deterministic predictability."

Finally, it is worth noting that randomness is not a state of "unpredictability," or "no information," as is sometimes thought. Rather, *random* means "not precisely predictable." For example, the amount of precipitation that will occur tomorrow where you live is a random quantity, not known to you today. However, a simple statistical analysis of climatological precipitation records at your location would yield relative frequencies of precipitation amounts that would provide substantially more information about tomorrow's precipitation at your location than I have as I sit writing this sentence. A still less uncertain idea of tomorrow's rain might be available to you in the form of a weather forecast. Reducing uncertainty about random meteorological events is the purpose of weather forecasts. Furthermore, statistical methods allow estimation of the precision of predictions, which can itself be valuable information.

1.4 An Aside on Notation

Specific values of a variable are conventionally distinguished by subscripts. For example, the quantities x_1, x_2, and x_3 would each represent different specific values of the variable x. These subscripts can be represented by variables known as *index variables*, which allow certain manipulations of the data to be expressed compactly using summation notation. Summation is indicated by the uppercase letter sigma, Σ. The particular members of a data set to be summed are indicated by notations above and below the sigma that pertain to the index variable. A familiar example of summation notation is the formula for the sample mean of n values of a set of data,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1.1}$$

The sample mean of a batch of numbers is simply obtained by adding up all the numbers in the batch and dividing by the number of values added. Here the summation is merely shorthand for the more cumbersome expression $x_1 + x_2 + x_3 + x_4 + \cdots + x_n$. While summation notation can look intimidating at first, remember that it is nothing more than an addition operator. The summation operator obeys all the usual rules of arithmetic, including, in particular, the distributive law of multiplication. The distributive law allows quantities that do not depend on the summation index to be factored out of, or multiplied into, the summation. In the case of the sample mean, for example,

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$
$$= \frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}. \tag{1.2}$$

A similar notation is conventionally used to indicate that a series of numbers is to be multiplied together. Here the symbol for the operator is a capital pi (for product):

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot x_3 \cdots \cdot x_n. \tag{1.3}$$

Chapter 2 Review of Probability

2.1 Background

The material in this chapter is meant to be a brief review of the basic elements of probability. More complete treatments of the basics of probability can be found in any good introductory statistics text, such as Dixon and Massey's (1983) *Introduction to Statistical Analysis* or Winkler's (1972) *Introduction to Bayesian Inference and Decision.*

Our uncertainty about the atmosphere, or about any other system, for that matter, is of different degrees in different instances. For example, you cannot be completely certain whether rain will occur at your home tomorrow, or whether the average temperature next month will be greater or less than the average temperature last month. But it is likely that you are more sure about the latter question than about the former one.

It is not sufficient, or even particularly informative, to say that a particular event is uncertain. Rather, one is faced with the problem of expressing or characterizing degrees of uncertainty. A possible approach is to use qualitative descriptors such as "likely," "unlikely," "possible," or "chance of." Conveying uncertainty through such phrases, however, is ambiguous and open to varying interpretations (Murphy and Brown, 1983). For example, it is not clear whether "rain likely" or "rain probable" indicates less uncertainty about the prospects for rain.

It is generally preferable to express uncertainty quantitatively, and this is done using numbers called *probabilities.* In a limited sense, probability is no more than an abstract mathematical system that can be developed logically from three premises called the *axioms of probability.* This system would be uninteresting to many people, including perhaps yourself, except that the resulting abstract concepts are relevant to real-world systems involving uncertainty. Before presenting the axioms of probability and a few of their more important implications, it is necessary to define some terminology.

2.2 The Elements of Probability

2.2.1 Events

An *event* is a set, or class, or group of possible uncertain outcomes. Events can be of two kinds. A *compound event* can be decomposed into two or more (sub)events, whereas an *elementary event* cannot. As a simple example, think about rolling an ordinary six-sided die. The event "an even number of spots comes up" is a compound event, since it will occur if either two, four, or six spots appear. The event "six spots come up" is an elementary event.

In simple situations like rolling dice it is usually obvious which events are simple and which are compound. But more generally, just what is defined to be elementary or compound often depends on the problem at hand and the purposes for which an analysis is being conducted. For example, the event "precipitation occurs tomorrow" could be an elementary event to be distinguished from the elementary event "precipitation does not occur tomorrow." But if it is important to distinguish further between forms of precipitation, "precipitation occurs" would be regarded as a compound event, possibly comprising the three elementary events "liquid precipitation," "frozen precipitation," and "both liquid and frozen precipitation." If one were interested further in how much precipitation will occur, these three events would themselves be regarded as compound, each composed of at least two elementary events. In this case, for example, the compound event "frozen precipitation" would occur if either of the elementary events "frozen precipitation containing at least 0.01-in. water equivalent" or "frozen precipitation containing less than 0.01-in. water equivalent" were to occur.

2.2.2 The Sample Space

The *sample space* or *event space* is the set of all possible elementary events. Thus the sample space represents the universe of all possible outcomes or events. Equivalently, it is the largest possible compound event.

The relationships among events in a sample space can be represented geometrically, using what is called a *Venn diagram*. Often the sample space is drawn as a rectangle and the events within it are drawn as circles, as in Fig. 2.1a. Here the sample space is the rectangle labeled S, which might be the set of possible precipitation outcomes for tomorrow. Four elementary events are depicted within the boundaries of the three circles. The "No precipitation" circle is drawn not overlapping the others because neither liquid nor frozen precipitation can occur if no precipitation occurs (i.e., in the absence of precipitation). The hatched area common to both "Liquid precipitation" and "Frozen precipitation" represents the event "both liquid and frozen precipitation." That part of S in Fig. 2.1a not surrounded by circles is interpreted as representing the "null event," which cannot occur.

Fig. 2.1 (a) A Venn diagram representing the relationships of selected precipitation events. The hatched region represents the event "both liquid and frozen precipitation." (b) An equivalent Venn diagram drawn to emphasize that S is composed of four MECE events.

It is not necessary to draw or think of Venn diagrams using circles to represent events. Figure 2.1b is an equivalent Venn diagram drawn using rectangles filling the entire sample space S. Drawn in this way, it is clear that S is composed of exactly four elementary events that represent the full range of outcomes that may occur. Such a collection of all possible elementary (according to whatever working definition is current) events is called *mutually exclusive and collectively exhaustive* (MECE). *Mutually exclusive* means that no more than one of the events can occur. *Collectively exhaustive* means that at least one of the events will occur. A set of MECE events completely fills a sample space.

Note that Fig. 2.1b could be modified to distinguish among precipitation amounts by adding a vertical line somewhere in the right-hand side of the rectangle. If the new rectangles on one side of this line were to represent precipitation of 0.01 in. or more, the rectangles on the other side would represent precipitation less than 0.01 in. The modified Venn diagram would then depict seven MECE events.

2.2.3 The Axioms of Probability

Having carefully defined the sample space and its constituent events, we must now associate probabilities with each event. The rules for doing so all flow logically from the three axioms of probability. While formal mathematical definitions are possible, the axioms can be stated somewhat loosely as

1. The probability of any event is nonnegative.
2. The probability of the compound event S is 1.
3. The probability of one or the other of two mutually exclusive events occurring is the sum of their two individual probabilities.

2.3 The Meaning of Probability

The axioms are the essential logical basis for the mathematics of probability. That is, the mathematical properties of probability can all be deduced from the axioms. A number of these properties are listed later in this chapter.

However, the axioms are not very informative about what probability actually means. There are two dominant views of the meaning of probability: the *frequency view* and the *Bayesian view*. Perhaps surprisingly, there has been no small controversy in the world of statistics as to which is correct. Passions have actually run so high on this issue that adherents to one view or the other have been known to launch personal (verbal) attacks on those in the opposite camp!

It is worth emphasizing that the mathematics is the same in either case. The differences are entirely in interpretation. Both of the interpretations of probability have been found to be accepted and useful in the atmospheric sciences, in much the same way that the particle–wave duality of the nature of electromagnetic radiation is accepted and useful in the field of physics.

2.3.1 Frequency Interpretation

This is the mainstream view of probability. Its development in the eighteenth century was motivated by the desire to understand games of chance. In this view, the probability of an event is exactly its long-run relative frequency. This definition is formalized in the *law of large numbers*, which states that the ratio of the number of occurrences of event $\{E\}$ to the number of opportunities for $\{E\}$ to have occurred converges to the probability of $\{E\}$, denoted $\Pr\{E\}$, as the number of opportunities increases. This idea can be written formally as

$$\Pr\left\{\left|\frac{a}{n} - \Pr\{E\}\right| \leq \epsilon\right\} \to 0 \qquad \text{as} \qquad n \to \infty, \tag{2.1}$$

where a is the number of occurrences, n is the number of opportunities (thus a/n is the relative frequency), and ϵ is an arbitrarily small number.

The frequency interpretation is intuitively reasonable and empirically sound. It is useful in such applications as estimating climatological probabilities by computing historical relative frequencies. For example, in the last 50 years there have been $31 \times 50 = 1550$ August days. If rain has occurred at a location of interest on 487 of those days, a natural estimate for the climatological probability of precipitation at that location during August would be $487/1550 = 0.314$.

2.3.2 Bayesian (Subjective) Interpretation

Strictly speaking, employing the frequency view of probability requires a long series of identical trials. For estimating climatological probabilities from his-

torical weather data this approach presents essentially no problem. However, thinking about probabilities for events such as {the football team at your college or alma mater will win at least one game next season} presents some difficulty in the relative frequency framework. While one can abstractly imagine a hypothetical series of football seasons identical to the upcoming one, this suite of fictitious football seasons is of no help in actually estimating a probability for the event.

The subjective interpretation is that probability represents the degree of belief, or quantified judgment, of a particular individual about the occurrence of an uncertain event. For example, there is now a long history of weather forecasters routinely (and very skillfully) assessing probabilities for events such as precipitation occurrence on days in the near future. If your college or alma mater is a large enough school that professional gamblers take an interest in the outcomes of its football games, probabilities regarding those outcomes are also regularly assessed—subjectively.

Two individuals can have different subjective probabilities for an event without either necessarily being wrong. This does not mean that an individual is free to choose any numbers and call them probabilities. The quantified judgment must be a consistent judgment. This means, among other things, that subjective probabilities must be consistent with the axioms of probability, and thus with the properties of probability implied by the axioms.

2.4 Some Properties of Probability

One reason Venn diagrams can be so useful is that they allow probabilities to be visualized geometrically as areas. One's familiarity with the geometric relationships in the physical world can then be used to better grasp the more ethereal world of probability. Imagine that the area of the rectangle in Fig. 2.1b is 1, according to the second axiom. The first axiom says that no areas can be negative. The third axiom says that the total area of nonoverlapping parts is the sum of the areas of those parts.

A number of mathematical properties of probability that follow logically from the axioms are listed in this section. The geometric analog for probability provided by a Venn diagram can be used to help visualize them.

2.4.1 Domain, Subsets, Complements, and Unions

Together, the first and second axiom imply that the probability of any event will be between zero and one, inclusive. This limitation on the domain of probability can be expressed mathematically as

$$0 \leq \Pr\{E\} \leq 1. \tag{2.2}$$

If $\Pr\{E\} = 0$ the event will not occur. If $\Pr\{E\} = 1$ the event is absolutely sure to occur.

If event $\{E_1\}$ necessarily occurs whenever event $\{E_2\}$ occurs, $\{E_1\}$ is said to be a subset of $\{E_2\}$. For example, $\{E_1\}$ and $\{E_2\}$ might denote occurrence of frozen precipitation, and occurrence of precipitation of any form, respectively. The third axiom implies

$$\Pr\{E_1\} \leq \Pr\{E_2\}. \tag{2.3}$$

The complement of event $\{E\}$ is the (generally compound) event that $\{E\}$ does not occur. In Fig. 2.1b, for example, the complement of the event "liquid and frozen precipitation" is the compound event "either no precipitation, liquid precipitation only, or frozen precipitation only." The complement of the event "at least 0.01 in. of liquid equivalent" is "either zero precipitation or less than 0.01 in. of precipitation." Together the second and third axioms imply

$$\Pr\{E\}^{\mathrm{C}} = 1 - \Pr\{E\}, \tag{2.4}$$

where $\{E\}^{\mathrm{C}}$ denotes the complement of $\{E\}$. Many authors use an overbar as an alternative notation to represent complements. This use of the overbar is very different from its most common statistical meaning, which is to denote an arithmetic average [Eq. (1.1)].

The union of two events is the compound event that one or the other, or both, of the events occur. In set notation, unions are denoted by the symbol $\cup$. As a consequence of the third axiom, probabilities for unions can be computed using

$$\begin{aligned}\Pr\{E_1 \cup E_2\} &= \Pr\{E_1 + E_2\} = \Pr\{E_1 \text{ or } E_2 \text{ or both}\} \\ &= \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \cap E_2\}.\end{aligned} \tag{2.5}$$

The symbol $\cap$ is called the the intersection operator, and

$$\Pr\{E_1 \cap E_2\} = \Pr\{E_1, E_2\} = \Pr\{E_1 \text{ and } E_2\} \tag{2.6}$$

is the event that both $\{E_1\}$ and $\{E_2\}$ occur. The notation $\{E_1, E_2\}$ is equivalent to $\{E_1 \cap E_2\}$. Another name for $\Pr\{E_1, E_2\}$ is the joint probability of $\{E_1\}$ and $\{E_2\}$. Equation (2.5) is sometimes called the *additive law of probability*. It holds regardless of whether $\{E_1\}$ and $\{E_2\}$ are mutually exclusive. However, if the two events are mutually exclusive, the probability of their intersection is zero, since mutually exclusive events cannot both occur.

The probability for the joint event, $\Pr\{E_1, E_2\}$ is subtracted in Eq. (2.5) to compensate for its having been counted twice when the probabilities for events $\{E_1\}$ and $\{E_2\}$ are added. This can be seen most easily by thinking about how to find the total geometric area surrounded by the two overlapping circles in Fig. 2.1a. The hatched region in Fig. 2.1a represents the intersection event {liquid precipitation and frozen precipitation}, and it is contained within each of the two circles labeled "Liquid precipitation" and "Frozen precipitation."

Fig. 2.2 Venn diagram illustrating computation of probability of the union of three intersecting events. The regions with two overlapping hatch patterns have been double-counted, and their areas must be subtracted to compensate. The central region with three overlapping hatch patterns has been triple-counted, but then subtracted three times when the double-counting was corrected. Its area must be added back again.

The additive law [Eq. (2.5)] can be extended to the union of three or more events by thinking of $\{E_1\}$ or $\{E_2\}$ as a compound event (i.e., a union of other events), and recursively applying Eq. (2.5). For example, if $\{E_2\} = \{E_3 \cup E_4\}$, substituting into Eq. (2.5) yields, after some rearrangement

$$\begin{aligned}\Pr\{E_1 \cup E_3 \cup E_4\} = {} & \Pr\{E_1\} + \Pr\{E_3\} + \Pr\{E_4\} \\ & - \Pr\{E_1 \cap E_3\} - \Pr\{E_1 \cap E_4\} - \Pr\{E_3 \cap E_4\} \\ & + \Pr\{E_1 \cap E_3 \cap E_4\}. \end{aligned} \tag{2.7}$$

This result may be difficult to grasp algebraically, but is fairly easy to visualize geometrically. Figure 2.2 illustrates the situation. Adding together the areas of the three circles individually [first line in Eq. (2.7)] results in double-counting of the areas with two overlapping hatch patterns, and triple-counting of the central area contained in all three circles. The second line of Eq. (2.7) corrects the double-counting, but subtracts the area of the central region three times. This area is added back a final time in the third line of Eq. (2.7).

2.4.2 Conditional Probability

It is often the case that one is interested in the probability of an event, given that some other event has occurred or will occur. For example, the probability of

freezing rain, given that precipitation occurs, may be of interest; or perhaps one needs to know the probability of coastal windspeeds above some threshold, given that a hurricane makes landfall nearby. These are examples of conditional probabilities. The event that must be "given" is called the *conditioning event.* The conventional notation for conditional probability is a vertical line, so denoting $\{E_1\}$ as the event of interest and $\{E_2\}$ as the conditioning event, one writes

$$\Pr\{E_1|E_2\} = \Pr\{E_1 \text{ given that } E_2 \text{ has occurred or will occur}\}. \quad (2.8)$$

If the event $\{E_2\}$ has occurred or will occur, the probability of $\{E_1\}$ is the conditional probability $\Pr\{E_1|E_2\}$. If the conditioning event has not occurred or will not occur, this in itself gives no information on the probability of $\{E_1\}$.

More formally, conditional probability is defined in terms of the intersection of the event of interest and the conditioning event, according to

$$\Pr\{E_1|E_2\} = \frac{\Pr\{E_1 \cap E_2\}}{\Pr\{E_2\}}, \quad (2.9)$$

provided the probability of the conditioning event is not zero. Intuitively, it makes sense that conditional probabilities are related to the joint probability of the two events in question, $\Pr\{E_1 \cap E_2\}$. Again, this is easiest to understand through the analogy to areas in a Venn diagram, as shown in Fig. 2.3. We understand the unconditional probability of $\{E_1\}$ to be represented by that proportion of the sample space S occupied by the rectangle labeled E_1. Conditioning on $\{E_2\}$ means that we are interested only in those outcomes containing $\{E_2\}$. We are, in effect, throwing away any part of S not contained in $\{E_2\}$. This amounts to considering a new sample space, S′, that is coincident with $\{E_2\}$. The conditional

Fig. 2.3 Illustration of the definition of conditional probability. The unconditional probability of $\{E_1\}$ is that fraction of the area of S occupied by $\{E_1\}$ on the left side of the figure. Conditioning on $\{E_2\}$ amounts to considering a new sample space, S′ composed only of $\{E_2\}$, since this means that we are concerned only with occasions when $\{E_2\}$ occurs. Therefore the conditional probability $\Pr\{E_1|E_2\}$ is given by that proportion of the area of the new sample space S′ occupied by both $\{E_1\}$ and $\{E_2\}$. This proportion is computed in Eq. (2.9).

probability $\Pr\{E_1|E_2\}$ is therefore represented geometrically as that proportion of the new sample space area occupied by both $\{E_1\}$ and $\{E_2\}$. If the conditioning event and the event of interest are mutually exclusive, the conditional probability clearly must be zero, since their joint probability will be zero.

2.4.3 Independence

Rearranging the definition of conditional probability Eq. (2.9) yields the form of this expression called the *multiplicative law of probability*:

$$\Pr\{E_1 \cap E_2\} = \Pr\{E_1|E_2\}\Pr\{E_2\} = \Pr\{E_2|E_1\}\Pr\{E_1\}. \tag{2.10}$$

Two events are said to be independent if the occurrence or nonoccurrence of one does not affect the probability of the other. For example, if one rolls a red die and a white die, the probability of an outcome of interest on the red die does not depend on the outcome of the white die, and vice versa. The outcomes of rolling dice are independent. Independence between $\{E_1\}$ and $\{E_2\}$ implies $\Pr\{E_1|E_2\} = \Pr\{E_1\}$ and $\Pr\{E_2|E_1\} = \Pr\{E_2\}$. Independence of events makes the calculation of joint probabilities particularly easy, since the multiplicative law then reduces to

$$\Pr\{E_1 \cap E_2\} = \Pr\{E_1\}\Pr\{E_2\}, \qquad \text{for } \{E_1\} \text{ and } \{E_2\} \text{ independent.} \tag{2.11}$$

Equation (2.11) is extended easily to the computation of joint probabilities for more than two independent events, by simply multiplying all the probabilities of the unconditional events.

Example 2.1. Conditional Relative Frequency

Consider estimating climatological (i.e., long-run, or relative frequency) estimates of probabilities using the data set given in Table A.1 of Appendix A. Climatological probabilities conditional on other events can be computed. Such probabilities are sometimes referred to as *conditional climatological probabilities*, or *conditional climatologies*.

Suppose it is of interest to estimate the probability of at least 0.01 in. of liquid equivalent precipitation at Ithaca in January, given that the minimum temperature is at least 0°F. Physically, these two events would be expected to be related since very cold temperatures typically occur on clear nights, and precipitation occurrence requires clouds. This physical relationship would lead one to expect that these two events would be statistically related (i.e., not independent), and that the conditional probabilities of precipitation given different temperature conditions will be different from each other, and from the unconditional probability. In particular, on the basis of our understanding of the underlying physical processes, we

expect the probability of precipitation given minimum temperature of 0° F or higher will be larger than the conditional probability given the complementary event of minimum temperature colder than 0° F.

To estimate this conditional relative frequency, one is interested only in those data records for which the Ithaca minimum temperature was at least 0° F. There are 24 such days in Table A.1. Of these 24 days, 14 show measurable precipitation (ppt), yielding the estimate $\Pr\{\text{ppt} \geq 0.01 \text{ in.} | T_{\min} \geq 0°\text{F}\} = 14/24 \approx 0.58$. The precipitation data for the 7 days on which the minimum temperature was colder than 0° F have been ignored. Since measurable precipitation was recorded on only one of these seven days, one could estimate the probability of precipitation given the complementary conditioning event of minimum temperature colder than 0° F as $\Pr\{\text{ppt} \geq 0.01 \text{ in.} | T_{\min} < 0°\text{F}\} = 1/7 \approx 0.14$. The corresponding estimate of the unconditional probability of precipitation would be $\Pr\{\text{ppt} \geq 0.01 \text{ in.}\} = 15/31 \approx 0.48$. ¤

The differences in the conditional probability estimates calculated in Example 2.1 reflect statistical dependence. Since the underlying physical processes are well understood, one would not be tempted to speculate that relatively warmer minimum temperatures somehow cause precipitation. Rather, the temperature and precipitation events show a statistical relationship because of their (different) physical relationships to clouds. When dealing with statistically dependent variables whose physical relationships may not be known, it is well to remember that statistical dependence does not necessarily imply a physical cause-and-effect relationship.

Example 2.2. Persistence as Conditional Probability

Atmospheric variables often exhibit statistical dependence with their own past or future values. In the terminology of the atmospheric sciences, this dependence through time is usually known as *persistence*. Persistence can be defined as the existence of (positive) statistical dependence among successive values of the same variable, or among successive occurrences of a given event. *Positive dependence* means that large values of the variable tend to be followed by relatively large values, and small values of the variable tend to be followed by relatively small values. It is usually the case that this statistical dependence of meteorological variables in time is positive. For example, the probability of an above-average temperature tomorrow is higher if today's temperature was above average. Thus, another name for persistence is *positive serial dependence*.

Consider characterizing the persistence of the event {precipitation occurrence} at Ithaca, again with the small data set in Table A.1 of Appendix A. Physically, serial dependence would be expected in this daily data because the typical time scale for the midlatitude synoptic waves associated with most winter precipitation at this location is several days. The statistical consequence should be that days for

which measurable precipitation is reported should tend to occur in runs, as should days without measurable precipitation.

In order to evaluate serial dependence for precipitation events it is necessary to estimate conditional probabilities of the type Pr{ppt today|ppt yesterday}. Since the data set in Table A.1 contains no records for either December 31, 1986 or February 1, 1987, there are 30 "yesterday/today" data pairs to work with. In order to estimate Pr{ppt today|ppt yesterday} one needs only to count the number of days reporting precipitation (as the conditioning, or "yesterday" event) that are followed by the subsequent day reporting precipitation (as the event of interest, or "today"). When estimating this conditional probability one is not interested in what happens following days on which no precipitation is reported. Excluding January 31, there are 14 days on which precipitation is reported. Of these, 10 are followed by another day with precipitation reported, and four are followed by dry days. The conditional relative frequency estimate would therefore be Pr{ppt today|ppt yesterday} = 10/14 ≈ 0.71. Similarly, conditioning on the complementary event (no precipitation "yesterday") yields Pr{ppt today|no ppt yesterday} = 5/16 ≈ 0.31. The difference in these conditional probability estimates confirms the serial dependence in these data, and quantifies the tendency of the wet and dry days to occur in runs. These two conditional probabilities also constitute a "conditional climatology." ¤

2.4.4 Law of total probability

Sometimes probabilities must be computed indirectly because of limited information. One relationship that can be useful in such situations is the *law of total probability*. Consider a set of MECE events, $\{E_i\}$, $i = 1, \ldots, I$ on the sample space of interest. Figure 2.4 illustrates this situation for $I = 5$ events. If there is an event $\{A\}$, also defined on this sample space, its probability can be computed by summing the joint probabilities

Fig. 2.4 Illustration of the law of total probability. The sample space S contains the event $\{A\}$, represented by the oval, and five MECE events, $\{E_i\}$.

$$\Pr\{A\} = \sum_{i=1}^{I} \Pr\{A \cap E_i\}. \tag{2.12}$$

Substituting the multiplicative law of probability yields

$$\Pr\{A\} = \sum_{i=1}^{I} \Pr\{A|E_i\}\Pr\{E_i\}. \tag{2.13}$$

If the unconditional probabilities $\Pr\{E_i\}$ and the conditional probabilities of $\{A\}$ given the MECE $\{E_i\}$ are known, the unconditional probability of $\{A\}$ can be computed. It is important to note that Eq. (2.13) is correct only if the events $\{E_i\}$ constitute a MECE partition of the sample space.

Example 2.3. Combining Conditional Probabilities Using the Law of Total Probability

Example 2.2 can also be viewed in terms of the law of total probability. Consider that there are only $I = 2$ MECE events partitioning the sample space: $\{E_1\}$ denotes precipitation yesterday and $\{E_2\} = \{E_1\}^C$ denotes no precipitation yesterday. Let event $\{A\}$ be the occurrence of precipitation today. If the data were not available, one could compute $\Pr\{A\}$ using the conditional probabilities through the law of total probability. That is, $\Pr\{A\} = \Pr\{A|E_1\}\Pr\{E_1\} + \Pr\{A|E_2\}\Pr\{E_2\} = (10/14)\,(14/30) + (5/16)\,(16/30) = 0.50$. Since the data are available in Appendix A, the correctness of this result can be confirmed simply by counting. ¤

2.4.5 Bayes' Theorem

Bayes' theorem is an interesting combination of the multiplicative law and the law of total probability. In a relative frequency setting, Bayes' theorem is used to "invert" conditional probabilities. That is, if $\Pr\{E_1|E_2\}$ is known, Bayes' theorem may be used to compute $\Pr\{E_2|E_1\}$. In the Bayesian framework it is used to revise or update subjective probabilities consistent with new information.

Consider again a situation such as that shown in Fig. 2.4, in which there are a defined set of MECE events $\{E_i\}$, and another event $\{A\}$. The multiplicative law [Eq. (2.10)] can be used to find two expressions for the joint probability of $\{A\}$ and any of the events $\{E_i\}$. Since

$$\begin{aligned}\Pr\{A, E_i\} &= \Pr\{A|E_i\}\Pr\{E_i\} \\ &= \Pr\{E_i|A\}\Pr\{A\},\end{aligned} \tag{2.14}$$

combining the two right-hand sides and rearranging yields

$$\Pr\{E_i|A\} = \frac{\Pr\{A|E_i\}\Pr\{E_i\}}{\Pr\{A\}} = \frac{\Pr\{A|E_i\}\Pr\{E_i\}}{\sum_{j=1}^{J} \Pr\{A|E_j\}\Pr\{E_j\}}. \tag{2.15}$$

The law of total probability has been used to rewrite the denominator. Equation (2.15) is the expression for Bayes' theorem. It is applicable separately for each of the MECE events $\{E_i\}$. Note, however, that the denominator is the same for each E_i, since $\Pr\{A\}$ is obtained each time by summing over all the events, indexed in the denominator by the subscript j.

Example 2.4. Bayes' Theorem from a Relative Frequency Standpoint

In Example 2.1, conditional probabilities for precipitation occurrence given minimum temperatures above or below 0° F were estimated. Bayes' theorem can be used to compute the converse conditional probabilities, concerning temperature events given that precipitation did or did not occur. Let $\{E_1\}$ represent minimum temperature of 0° F or above, and let $\{E_2\} = \{E_1\}^C$ be the complementary event that minimum temperature is colder than 0° F. Clearly the two events constitute a MECE partition of the sample space. Recall that minimum temperatures of at least 0° F were reported on 24 of the 31 days, so that the unconditional climatological estimates of probabilities for the temperature events would be $\Pr\{E_1\} = 24/31$ and $\Pr\{E_2\} = 7/31$. Recall from Example 2.1 that $\Pr\{A|E_1\} = 14/24$ and $\Pr\{A|E_2\} = 1/7$.

Equation 2.15 can be applied separately for each of the two events $\{E_i\}$. In each case the denominator is $\Pr\{A\} = (14/24)\,(24/31) + (1/7)\,(7/31) = 15/31$. (This differs slightly from the estimate for the probability of precipitation derived in Example 2.2, since there the data for December 31 could not be included.) Using Bayes' theorem, the conditional probability for minimum temperature at least 0° F given precipitation occurrence is $(14/24)\,(24/31)/(15/31) = 14/15$. Similarly, the conditional probability for minimum temperature below 0° F given nonzero precipitation is $(1/7)\,(7/31)/(15/31) = 1/15$. Since all the data are available in Appendix A, these calculations can also be verified directly. ¤

Example 2.5. Bayes' Theorem from a Subjective Probability Standpoint

A subjective (Bayesian) probability interpretation of Example 2.4 can also be made. Suppose a weather forecast specifying the probability of the minimum temperature being at least 0° F is desired. If no more sophisticated information were available, it would be natural to use the unconditional climatological probability for the event, $\Pr\{E_1\} = 24/31$, as representing the forecaster's uncertainty or degree of belief in the outcome. In the Bayesian framework this baseline state of information is known as the *prior probability*. Suppose, however, that the forecaster could know whether precipitation will occur on that day. This information will affect the forecaster's degree of certainty in the temperature outcome. Just how much more certain the forecaster can become depends on the strength of the relationship between temperature and precipitation, expressed in the conditional probabilities of precipitation occurrence given the two minimum temperature outcomes. These conditional probabilities, $\Pr\{A|E_i\}$ in the notation of this example,

are known as the *likelihoods*. If precipitation occurs, the forecaster is more certain that the minimum temperature will be at least 0°F, with the revised probability given by Eq. (2.15) as (14/24) (24/31)/(15/31) = 14/15. This modified, or updated, judgment regarding the probability of a very cold minimum temperature not occurring is called the *posterior probability*. Here the posterior probability is substantially larger than the prior probability of 24/31. Similarly, if precipitation does not occur, the forecaster is more confident that the minimum temperature will not be 0°F or warmer. Note that the differences between this example and Example 2.4 are entirely in the interpretation, and that the computations and numerical results are identical. ¤

Exercises

2.1. In the climatic record for 60 winters at a given location, single-storm snowfalls greater than 35 cm occurred in nine of those winters (define such snowfalls as event A), the lowest temperature was below $-25°$C in 36 of the winters (define this as event B). Both events A and B occurred in three of the winters.

(a) Sketch a Venn diagram for a sample space appropriate to these data.

(b) Write an expression using set notation, for the occurrence of 35-cm snowfalls, $-25°$C temperatures, or both. Estimate the climatological probability for this compound event.

(c) Write an expression using set notation, for the occurrence of winters with 35-cm snowfalls in which the temperature does not fall below $-25°$C. Estimate the climatological probability for this compound event.

(d) Write an expression using set notation, for the occurrence of winters having neither $-25°$C temperatures nor 35-cm snowfalls. Again, estimate the climatological probability.

2.2. Using the January 1987 data set in Table A.1, define event A as Ithaca $T_{max} > 32°$F and event B as Canandaigua $T_{max} > 32°$F.

(a) Explain the meanings of $\Pr(A)$, $\Pr(B)$, $\Pr(A,B)$, $\Pr(A \cup B)$, $\Pr(A|B)$, and $\Pr(B|A)$.

(b) Estimate, using relative frequencies in the data, $\Pr(A)$, $\Pr(B)$, and $\Pr(A,B)$.

(c) Using the results from part (b), calculate $\Pr(A|B)$.

(d) Are events A and B independent? How do you know?

2.3. Three radar sets, operating independently, are searching for "hook" echos (a radar signature associated with tornados). Suppose that each radar has a probability of 0.05 of failing to detect this signature when a tornado is present.

(a) Sketch a Venn diagram for a sample space appropriate to this problem.

(b) What is the probability that a tornado will escape detection by all three radars?

(c) What is the probability that a tornado will be detected by all three radars?

2.4. The effect of cloud seeding on suppression of damaging hail is being studied in your area, by randomly seeding or not seeding equal numbers of candidate storms. Suppose the probability of damaging hail from a seeded storm is 0.10, and the probability of damaging hail from an unseeded storm is 0.40. If one of the candidate storms has just produced damaging hail, what is the probability that it was seeded?

Chapter 3 Empirical Distributions and Exploratory Data Analysis

3.1 Background

One very important application of statistical ideas in meteorology and climatology is in making sense of a new set of data. As mentioned in Chapter 1, meteorological observing systems and numerical models, supporting both operational and research efforts, produce torrents of numerical data. It can be a significant task just to get a feel for a new batch of numbers, and to begin to make some sense of them. The goal is to extract insight about the natural processes underlying the generation of the numbers.

Broadly speaking, this activity is known as *exploratory data analysis* (EDA). Its systematic use increased substantially following Tukey's (1977) pathbreaking and very readable book of the same name. The methods of EDA draw heavily on a variety of graphical methods to aid in the comprehension of the sea of numbers confronting an analyst. Graphics are a very effective means of compressing and summarizing data, portraying much in little space, and exposing unusual features of a data set. The unusual features can be especially important. Sometimes unusual data points result from errors in recording or transcription, and it is well to know about these as early as possible in an analysis. Sometimes the unusual data are valid, and may turn out be the most interesting and informative parts of the data set.

Many EDA methods were originally designed to be applied by hand, with pencil and paper, to small (up to perhaps 200-point) data sets. More recently, graphically oriented computer packages have come into being that allow fast and easy use of these methods on desktop computers (e.g., Velleman, 1988). The methods can also be implemented on larger computers with a small amount of programming in, for example, the FORTRAN or C languages.

3.1.1 Robustness and Resistance

Most of the classical techniques of statistics work best when fairly stringent assumptions about the nature of the data are met. For example, it is often assumed

that data will follow the familiar bell-shaped curve of the Gaussian distribution. The classical procedures can behave very badly, indeed (i.e., produce quite misleading results) if their assumptions are not satisfied by the data at hand.

Historically, the assumptions of classical statistics were made not out of ignorance but rather out of necessity. Invocation of simplifying assumptions in statistics, as in other fields, has allowed progress to be made through the derivation of elegant analytic results, which are relatively simple but powerful mathematical formulas. As has been the case in many quantitative fields, the relatively recent advent of cheap computing power has freed the data analyst from sole dependence on such results, by allowing less stringent alternatives to become practical. This does not mean that the classical methods are no longer useful. However, it is much easier to check that a given set of data satisfies particular assumptions before a classical procedure is used.

Two important properties of EDA methods are that they are *robust* and *resistant*. Robustness and resistance are two aspects of insensitivity to assumptions about the nature of a set of data. A robust method is not necessarily optimal in any particular circumstance, but performs reasonably well in most circumstances. For example, the sample average [Eq. (1.1)] is the best characterization of the center of a set of data if it is known that those data follow a Gaussian distribution. However, if those data are decidedly non-Gaussian (e.g., if they are a record of extreme rainfall events), the sample average will yield a misleading characterization of their center. In contrast, robust methods generally are not sensitive to particular assumptions about the overall nature of the data.

A resistant method is not unduly influenced by a small number of outliers, or "wild data." As indicated previously, such points often show up in a batch of data through errors of one kind or another. The results of a resistant method change very little if a small fraction of the data values are changed, even if they are changed drastically. In addition to not being robust, the sample average is not a resistant characterization of the center of a data set, either. Consider the small set $\{11, 12, 13, 14, 15, 16, 17, 18, 19\}$. Its average is 15. However, if a transcription error had resulted instead in the set $\{11, 12, 13, 14, 15, 16, 17, 18, 91\}$, the "center" of the data (erroneously) characterized using the sample average would instead be 23. Resistant measures of the center of a batch of data, such as those to be presented below, would be changed little or not at all by the substitution of "91" for "19" in this simple example.

3.1.2 Quantiles

Many common summary measures rely on the use of selected sample *quantiles*, also known as *fractiles*. These terms are essentially equivalent to the more familiar term *percentile*. A sample quantile q_p is a number having the same units as the data, which exceeds that proportion of the data given by the subscript p, with $0 \leq$

$p \leq 1$. The sample quantile q_p can be interpreted approximately as that data value exceeding a randomly chosen member of the data set, with probability p. Equivalently, the sample quantile q_p would be regarded as the $[p \times 100]$th percentile of the data set.

The determination of sample quantiles requires that a batch of data first be arranged in order. Sorting small sets of data by hand presents little problem. Sorting larger sets of data is best accomplished by computer. Historically, the sorting step constituted a major bottleneck in the application of robust and resistant procedures to large data sets. Today the sorting can be easily done using either a spreadsheet or data analysis program on a desktop computer, or one of many sorting algorithms available in collections of general-purpose computing routines (e.g., Press *et al.*, 1986).

The sorted, or ranked, data values from a particular sample are called the *order statistics* of that sample. Given a set of data $\{x_1, x_2, x_3, x_4, x_5, \ldots, x_n\}$, the order statistics of the sample would be the same numbers, sorted in ascending order. These sorted values are conventionally denoted using parenthetical subscripts, i.e., by the set $\{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, \ldots, x_{(n)}\}$. Here the ith smallest of the n data values is denoted $x_{(i)}$.

Certain sample quantiles are used particularly often in the exploratory summarization of data. Most commonly used is the *median*, or $q_{0.5}$, or 50th percentile. This is the value at the center of the data set, in the sense that equal proportions of the data fall above and below it. If the data set at hand contains an odd number of values, the median is simply the middle order statistic. If there are an even number, however, the data set has two middle values. In this case the median is conventionally taken to be the average of these two middle values. Formally,

$$q_{0.5} = \begin{cases} x_{([n+1]/2)}, & n \text{ odd} \\ \dfrac{x_{(n/2)} + x_{([n/2]+1)}}{2}, & n \text{ even} \end{cases} \tag{3.1}$$

Almost as commonly used as the median are the *quartiles*, $q_{0.25}$ and $q_{0.75}$. Usually these are called the *lower* and *upper* quartiles, respectively. They are locate d halfway between the median, $q_{0.5}$, and the extremes, $x_{(1)}$ and $x_{(n)}$. In typically colorful terminology, Tukey (1977) calls $q_{0.25}$ and $q_{0.75}$ the "hinges," imagining that the data set has been folded first at the median, and then at the quartiles. The quartiles are thus the two medians of the half–data sets between $q_{0.5}$ and the extremes. If n is odd, these half–data sets each consist of $(n + 1)/2$ points, and both include the median. If n is even, these half–data sets each contain $n/2$ points, and do not overlap. Less commonly used quantiles are the *eighths*, $q_{0.125}$, $q_{0.325}$, $q_{0.625}$, and $q_{0.825}$; and the *sixteenths*, $q_{0.0625}$, and so on. Also used are the *deciles*, $q_{0.1}, q_{0.2}, \ldots, q_{0.9}$.

Example 3.1. Computation of Common Quantiles

If there are $n = 9$ data values in a batch of data, the median is $q_{0.5} = x_{(5)}$, or the fifth largest of the nine. The lower quartile is $q_{0.25} = x_{(3)}$, and the upper quartile is $q_{0.75} = x_{(7)}$.

If $n = 10$, the median is the average of the two middle values, and the quartiles are the single middle values of the upper and lower halves of the data. That is, $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ are $x_{(3)}$, $[x_{(5)} + x_{(6)}]/2$, and $x_{(8)}$, respectively

If $n = 11$, then there is a unique middle value, but the quartiles are determined by averaging the two middle values of the upper and lower halves of the data. That is, $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ are $[x_{(3)} + x_{(4)}]/2$, $x_{(6)}$, and $[x_{(8)} + x_{(9)}]/2$, respectively.

For $n = 12$ both quartiles and the median are determined by averaging pairs of middle values; $q_{0.25}$, $q_{0.5}$, and $q_{0.75}$ are $[x_{(3)} + x_{(4)}]/2$, $[x_{(6)} + x_{(7)}]/2$, and $[x_{(9)} + x_{(10)}]/2$, respectively. ¤

3.2 Numerical Summary Measures

Some simple robust and resistant summary measures are available that can be used without hand plotting or computer graphic capabilities. Often these will be the first quantities to be computed from a new and unfamiliar set of data. The numerical summaries listed below can be subdivided into measures of location, spread, and symmetry. *Location* refers to the central tendency, or general magnitude, of the data values. *Spread* denotes the degree of variation or dispersion around the central value. *Symmetry* describes the degree to which the data values are evenly distributed about their center. Asymmetric data tend to spread more either on the high side (have a long right tail), or the low side (have a long left tail). These three types of numerical summary measures correspond to the first three statistical moments of a data sample, but the usual measures of these moments (i.e., the sample mean, sample variance, and sample coefficient of skewness, respectively) are neither robust nor resistant.

3.2.1 Location

The most common robust and resistant measure of central tendency is the median, $q_{0.5}$. Consider again the data set {11, 12, 13, 14, 15, 16, 17, 18, 19}. The median and mean are both 15. If, as noted before, the "19" is replaced erroneously by "91," the mean is very strongly affected. The median is unchanged by this common type of data error.

A slightly more complicated measure of location that takes into account more information about the magnitudes of the data is the trimean. The *trimean* is a weighted average of the median and the quartiles, with the median receiving twice the weight of each of the quartiles:

$$\text{Trimean} = \frac{q_{0.25} + 2\,q_{0.5} + q_{0.75}}{4}. \tag{3.2}$$

The trimmed mean is another resistant measure of location, whose sensitivity to outliers is reduced by removing a specified proportion of the largest and smallest observations. If the proportion of observations omitted at each end is α, then the α-trimmed mean is

$$\bar{x}_\alpha = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x_{(i)}, \tag{3.3}$$

where k is an integer rounding of the product αn, the number of data values "trimmed" from each tail. The trimmed mean reduces to the ordinary mean [Eq. (1.1)] for $\alpha = 0$.

Other methods of characterizing location can be found in Andrews *et al.* (1972), Goodall (1983), Rosenberger and Gasko (1983), and Tukey (1977).

3.2.2 Spread

The most common, and simplest, robust and resistant measure of spread, also known as *dispersion*, or *scale*, is the *interquartile range* (IQR). The interquartile range is simply the difference between the upper and lower quartiles:

$$\text{IQR} = q_{0.75} - q_{0.25}. \tag{3.4}$$

The IQR is a good index of the spread of a data set in the central part of the data, since it simply specifies the range of the central 50% of the data. The fact that it ignores the upper and lower 25% of the data makes it quite resistant to outliers. This quantity is sometimes called the *fourth-spread.*

It is worthwhile to compare the IQR with the conventional measure of scale of a data set, the sample *standard deviation*

$$s = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{3.5}$$

The square of the sample standard deviation, s^2, is known as the *sample variance.* The standard deviation is neither robust nor resistant. It is very nearly just the square root of the average squared difference between the data points and their sample mean. Because of the square root in Eq. (3.5), the standard deviation has the same physical dimensions as the underlying data. Even one very large data value will be felt very strongly because it will be especially far away from the mean, and that difference will be magnified by the squaring process. Considering again the set {11, 12, 13, 14, 15, 16, 17, 18, 19}, the sample standard deviation is 2.74, but is greatly inflated to 25.6 if "91" erroneously replaces "19." It is easy to see that in either case IQR = 4.

While the IQR is very easy to compute, it does have the disadvantage of not making much use of much of the data. A more complete, yet reasonably simple alternative is the *median absolute deviation* (MAD). The MAD is easiest to understand by imagining the transformation $y_i = |x_i - q_{0.5}|$. Each transformed value y_i is the absolute value of the difference between the original data values and their median. The MAD is then just the median of the transformed (y_i) values:

$$\text{MAD} = \text{median}|x_i - q_{0.5}|. \tag{3.6}$$

A still more elaborate measure of spread is the *trimmed variance*. The idea, as for the trimmed mean [Eq. (3.3)], is to omit a proportion of the largest and smallest values and compute the analog of the sample variance [the square of Eq. (3.5)]

$$s_\alpha^2 = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} (x_{(i)} - \bar{x}_\alpha)^2. \tag{3.7}$$

Again, k is the nearest integer to αn, and squared deviations from the consistent trimmed mean [Eq. (3.3)] are averaged. The trimmed variance is sometimes multiplied by an adjustment factor to make it more consistent with the ordinary sample variance, s^2 (Graedel and Kleiner, 1985).

Other measures of spread can be found in Hosking (1990) and Iglewicz (1983).

3.2.3 Symmetry

The conventional moments-based measure of symmetry in a batch of data is the *sample skewness coefficient*

$$\gamma = \frac{[1/(n - 1)] \sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}. \tag{3.8}$$

This measure is neither robust nor resistant. The numerator is similar to the sample variance, except that the average is over cubed deviations from the mean. Thus the sample skewness coefficient is even more sensitive to outliers than is the standard deviation. The average cubed deviation in the numerator is divided by the cube of the sample standard deviation in order to standardize the skewness coefficient and make comparisons of skewness between different data samples more meaningful. The standardization also serves to make the sample skewness a dimensionless quantity.

Notice that cubing differences between the data values and their mean preserves the signs of these differences. Since the differences are cubed, the data values farthest from the mean will dominate the sum in the numerator of Eq. (3.8). If there are a few very large data values, the sample skewness will tend to be positive. Therefore batches of data with long right tails are referred to both as *right-skewed* and *positively skewed*. Data that are physically constrained to lie above a minimum value (such as precipitation or windspeed, which must be nonnegative)

are often positively skewed. Conversely, if there are a few very small (or large negative) data values, these will fall far below the mean. The sum in Eq. (3.8) will then be dominated by a few large negative terms, so that the skewness coefficient will be negative. Data with long left tails are referred to as *left-skewed*, or *negatively skewed*. For essentially symmetric data, the skewness coefficient will be near zero.

A robust and resistant alternative to the sample skewness is the *Yule–Kendall index*

$$\gamma_{\mathrm{YK}} = \frac{(q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25})}{\mathrm{IQR}} = \frac{q_{0.25} - 2\,q_{0.5} + q_{0.75}}{\mathrm{IQR}}, \quad (3.9)$$

which is computed by comparing the distance between the median and each of the two quartiles. If the data are right-skewed, at least in the central 50% of the data, the distance to the median will be greater from the upper quartile than from the lower quartile. In this case the Yule–Kendall index will be greater than zero, consistent with the usual convention of right-skewness being positive. Conversely, left-skewed data will be characterized by a negative Yule–Kendall index. Analogously to Eq. (3.8), division by the interquartile range nondimensionalizes γ_{YK} (i.e., scales it in a way that the physical dimensions, such as meters or millibars, cancel) and thus improves its comparability between data sets.

Alternative measures of skewness can be found in Brooks and Carruthers (1953) and Hosking (1990).

3.3 Graphical Summary Techniques

While numerical summary measures are quick and easy to compute and display, they can express only a small amount of detail. In addition, their visual impact is limited. A number of graphical displays for exploratory data analysis have been devised that require only slightly more effort to produce.

3.3.1 Stem-and-Leaf Display

The *stem-and-leaf display* is a very simple but effective tool for producing an overall view of a new set of data. At the same time it provides the analyst with an initial exposure to the individual data values. In its simplest form, the stem-and-leaf display groups the data values according to their all-but-least-significant digits. These values are written in either ascending or descending order to the left of a vertical bar, and these constitute the "stems." The least-significant digit for each data value is then written to the right of the vertical bar, on the same line as the more significant digits with which it belongs. These least-significant values constitute the "leaves."

(a)

```
5 | 3
4 | 5
3 | 3 2 0 0 7 7 0 6 2 3 4 2 0 4
2 | 9 5 9 5 8 7 6 8 4 6 2 6 7
1 | 7
0 | 9
```

(b)

```
5• | 3
4* | 5
4• |
3* | 6 7 7
3• | 0 0 0 0 2 2 2 3 3 4 4
2* | 5 5 6 6 6 7 7 8 8 9 9
2• | 2 4
1* | 7
1• |
0* | 9
```

Fig. 3.1 Stem-and-leaf displays for the January 1987 Ithaca maximum temperatures in Table A.1. The plot in panel a results after the first pass through the data, using the 10s as "stem" values. A bit more resolution is obtained in panel b by creating separate stems for least-significant digits from 0 to 4 (•), and from 5 to 9 (*). At this stage it is also easy to sort the data values before rewriting them.

Figure 3.1a shows a stem-and-leaf display for the January 1987 Ithaca maximum temperatures in Table A.1. These data values are reported to whole degrees, and range from 9 to 53°F. The all-but-least-significant digits are thus the tens of degrees, which are written to the left of the bar. The display is built up by proceeding through the data values one by one, and writing its least-significant digit on the appropriate line. For example, the temperature for January 1 is 33°F, so the first leaf to be plotted is the first "3" on the stem of temperatures in the 30s. The temperature for January 2 is 32°F, so a "2" is written to the right of the "3" just plotted for January 1.

The initial stem-and-leaf display for this particular data set is a bit crowded, since most of the values in this particular data set are in the 20s and 30s. More resolution can be obtained by constructing a second plot, like that in Fig. 3.1b, where each stem has been split to contain only the values 0–4 or 5–9. Sometimes the opposite problem will occur, and the initial plot is too sparse. In that case (if there are at least three significant digits), replotting can be done with stem labels omitting the two least-significant digits. Less stringent groupings can also be used. Regardless of whether it may be desirable to split or consolidate stems, it is often useful to rewrite the display with the leaf values sorted, as has also been done in Fig. 3.1b.

The stem-and-leaf display is much like a quickly plotted histogram of the data, placed on its side. In Fig. 3.1, for example, it is evident that these temperature data are reasonably symmetrical, with most of the values falling in the upper 20s and lower 30s. Sorting the leaf values also facilitates extraction of quantiles of interest. In this case it is easy to count inward from the extremes to find that the median is 30, and that the two quartiles are 26 and 33.

It can happen that there are one or more outlying data points that are far removed from the main body of the set. Rather than plot many empty stems, it is usually more convenient to just list these values separately at the upper and lower

```
High: 38.8, 51.9

    25 | 9
    24 | 0
    23 |
    22 |
    21 |
    20 |
    19 |
    18 | 55
    17 |
    16 | 6666
    15 |
    14 |
    13 |
    12 | 9999
    11 | 1111111
    10 |
     9 | 22222
     8 |
     7 | 4444
Low: 0.0, 0.0
```

Fig. 3.2 Stem-and-leaf display of 1:00 A.M. windspeeds (km/hr) at Newark, New Jersey, Airport during December 1974. Very high and very low values are written outside the plot itself to avoid having many blank stems. The striking grouping of the leaf values suggests that a roundoff process has been applied to the original observations. [After Graedel and Kleiner (1985).]

ends of the display, as in Fig. 3.2. This display is of data, taken from Graedel and Kleiner (1985), of windspeeds in kilometers per hour (km/h) to the nearest tenth. Merely listing two extremely large values and two values of calm winds at the top and bottom of the plot has reduced the length of the display by half. It is evident that the data are strongly skewed to the right, as often occurs for wind data.

The stem-and-leaf display in Fig. 3.2 also reveals something that might have been missed in a tabular list of the daily data. All the leaf values on each stem are the same. Evidently a rounding process has been applied to the data, knowledge of which could be important to some subsequent analyses. In this case the rounding process consists of transforming the data from the original units (knots) to kilometers per hour. For example, the four observations of 16.6 km/h result from original observations of 9 knots. No observations on the 17-km/h stem would be possible, since observations of 10 knots transform to 18.5 km/h.

3.3.2 Boxplots

The *boxplot*, or *box-and-whisker plot*, is a very widely used graphical tool introduced by Tukey (1977). It is a simple plot of five sample quantiles: the minimum, $x_{(1)}$; the lower quartile, $q_{0.25}$; the median, $q_{0.5}$; the upper quartile, $q_{0.75}$; and the maximum, $x_{(n)}$. Using these five numbers, the boxplot essentially presents a quick sketch of the distribution of the underlying data.

Fig. 3.3 A simple boxplot, or box-and-whiskers plot, for the January 1987 Ithaca maximum temperature data. The upper and lower ends of the box are drawn at the quartiles, and the bar through the box is drawn at the median. The whiskers extend from the quartiles to the maximum and minimum data values.

Figure 3.3 shows a boxplot for the January 1987 Ithaca maximum temperature data in Table A.1. The box in the middle of the diagram is bounded by the upper and lower quartiles, and thus locates the central 50% of the data. The bar inside the box locates the median. The whiskers extend away from the box to the two extreme values.

Boxplots can convey a surprisingly large amount of information at a glance. It is clear from the small range occupied by the box in Fig. 3.3, for example, that the data are concentrated quite near 30° F. Being based only on the median and the quartiles, this portion of the boxplot is highly resistant to any outliers that might be present. The full range of the data is also apparent at a glance. Finally, it is apparent that these data are nearly symmetrical, since the median is near the center of the box, and the whiskers are of comparable length.

3.3.3 Schematic Plots

A shortcoming of the boxplot is that information about the tails of the data is highly generalized. The whiskers extend to the highest and lowest values, but there is no information about the distribution of data points within the upper and lower quarters of the data. For example, Fig. 3.3 shows that the highest maximum temperature is 53° F but gives no information as to whether this is an isolated point (with the remaining warm temperatures cooler than, say, 40° F) or whether the warmer temperatures are more-or-less evenly distributed between the upper quartile and the maximum.

It is often useful to have some idea of the degree of unusualness of the extreme values. A refinement on the boxplot that presents more detail in the tails is the *schematic plot*, which was also originated by Tukey (1977). The schematic plot is identical to the boxplot, except that extreme points deemed to be sufficiently un-

usual are plotted individually. Just how extreme is sufficiently unusual depends on the intrinsic variability of the data in the central part of the sample, as reflected by the quartiles. A given extreme value is regarded as being less unusual if the two quartiles are far apart (i.e., if the IQR is large), and more unusual if the two quartiles are very near each other (the IQR is small).

The dividing lines between less- and more-unusual points are known in Tukey's idiosyncratic terminology as the "fences." Four fences are defined: inner and outer fences, above and below the data, according to

$$\begin{aligned}
\text{Upper outer fence} &= q_{0.75} + 3\ \text{IQR}, \\
\text{Upper inner fence} &= q_{0.75} + \frac{3\ \text{IQR}}{2}, \\
\text{Lower inner fence} &= q_{0.25} - \frac{3\ \text{IQR}}{2}, \\
\text{Lower outer fence} &= q_{0.25} - 3\ \text{IQR}.
\end{aligned} \tag{3.10}$$

Thus the two outer fences are located 3 times the distance of the interquartile range above and below the two quartiles. The inner fences are midway between the outer fences and the quartiles, being 1½ times the distance of the interquartile range away from the quartiles.

In the schematic plot, points within the inner fences are called "inside." The range of the inside points is shown by the extent of the whiskers. Data points between the inner and outer fences are referred to as being "outside," and are plotted individually in the schematic plot. Points above the upper outer fence or below the lower outer fence are called "far out," and are plotted individually with a different symbol. In common with the boxplot, the box in a schematic plot shows the locations of the quartiles and the median.

Example 3.2. Construction of a Schematic Plot

Figure 3.4 is a schematic plot for the January 1987 Ithaca maximum temperature data. As can be determined from Fig. 3.1, the quartiles for this data are at 33 and 26°F, and the IQR = 33 − 26 = 7°F. From this information it is easy to compute the locations of the inner fences at 33 + (3/2)(7) = 43.5°F and 26 − (3/2)(7) = 15.5°F. Similarly, the outer fences are at 33 + (3)(7) = 54°F and 26 − (3)(7) = 5°F.

The two warmest temperatures, 53 and 45°F, are greater than the upper inner fence, and are shown individually by circles. The coldest temperature, 9°F, is less than the lower inner fence, and is also plotted individually. The whiskers are drawn to the most extreme temperatures inside the fences, 37 and 17°F. If the warmest temperature had been 55°F, rather than 53°F, it would fall outside the outer fence ("far out"), and would have been plotted individually with a different symbol. This separate symbol for the "far out" points is often an asterisk. ¤

Fig. 3.4 A schematic plot for the January 1987 Ithaca maximum temperature data. The central "box" portion of the figure is identical to the boxplot of the same data in Fig. 3.3. The three values "outside" the inner fences are plotted separately. None of the values are beyond the outer fences, or "far out." Notice that the whiskers extend to the most extreme inside data values, and not to the fences.

One important use of schematic plots or box plots is simultaneous graphical comparison of several batches of data. This use of schematic plots is illustrated in Fig. 3.5, which shows side-by-side plots for all four of the batches of temperature data in Table A.1. Of course, it is known in advance that the maximum temperatures are warmer than the minimum temperatures, and comparing their schematic

Fig. 3.5 Side-by-side schematic plots for the January 1987 temperatures in Table A.1. The minimum temperature data for both locations are all "inside," so the schematic plots are identical to ordinary boxplots.

plots brings out this difference quite strongly. Apparently, Canandaigua was slightly warmer than Ithaca during this month, and more strongly so for the minimum temperatures. The Ithaca minimum temperatures were evidently more variable than the Canandaigua minimum temperatures. For both locations, the minimum temperatures are more variable than the maximum temperatures, especially in the central parts of the distributions represented by the boxes. The location of the median in the upper end of the boxes of the minimum temperature schematic plots indicates a tendency toward negative skewness, as does the inequality of the whisker lengths for the Ithaca data. The maximum temperatures appear to be reasonably symmetrical for both locations. Note that none of the minimum temperature data are outside the inner fences, so that boxplots of the same data would be identical.

3.3.4 Other Boxplot Variants

Two variations on boxplots or schematic plots suggested by McGill *et al.* (1978) are sometimes used, particularly when comparing side-by-side plots. The first is to plot each box width proportional to $\sqrt{n}$. This simple variation allows plots of data having larger sample sizes to stand out and give a stronger visual impact.

The second variant is the *notched* boxplot or schematic plot. The boxes in these plots resemble hourglasses, with the constriction or waist located at the median. The lengths of the notched portions of the box differ from plot to plot, reflecting estimates of preselected confidence limits (see Chapter 5) for the median. The details of constructing these intervals are given in Velleman and Hoaglin (1981). Combining both of these techniques, i.e., constructing notched, variable-width plots, is straightforward. If the notched portion needs to extend beyond the quartiles, however, the overall appearance of the plot can begin to look a bit strange [an example can be seen in Graedel and Kleiner (1985)]. A nice alternative to notching is to add shading or stippling in the box to span the computed interval, rather than deforming its outline with notches (e.g., Velleman, 1988).

3.3.5 Histograms

The *histogram* is a very familiar graphical display device for a single batch of data. The range of the data is divided into class intervals, or "bins," and the number of values falling into each interval is counted. The histogram then consists of a series of rectangles whose widths are defined by the class limits, and whose heights depend on the number of values in each interval. Example histograms are shown in Figs. 4.1 and 4.14. Histograms quickly reveal such attributes of the data as location, spread, and symmetry. If the data are multimodal (i.e., more than one "hump" in the distribution of the data), this is quickly evident as well.

Most commonly, the widths of the class intervals are chosen to be equal. In this case the heights of the histogram bars are then simply proportional to the numbers

of counts. The vertical axis can be labeled to give either the number of counts represented by each bar (the absolute frequency) or the proportion of the entire sample represented by each bar (the relative frequency). More properly, however, it is the areas of the histogram bars that are proportional to the number of counts. This point becomes important if the class intervals for the histogram bars are chosen to have unequal widths, and when a theoretical probability function is to be superimposed on the histogram.

The main issue to be confronted when constructing a histogram is choice of the interval width. Intervals that are too wide will result in important details of the data being masked. Intervals that are too narrow will result in a plot that is irregular and difficult to interpret. Some guidance for the class width, w, of equal-interval histograms can be taken from the results collected by Emerson and Hoaglin (1983). Two such suggestions are

$$w \cong \mathrm{IQR}\left[\frac{\ln(n)}{n}\right]^{1/3} \tag{3.11a}$$

and

$$w \cong \frac{c\ \mathrm{IQR}}{n^{1/3}}, \tag{3.11b}$$

where c is a constant in the range of perhaps 2.0–2.5. Approaching the problem from the opposite perspective, Panofsky and Brier (1958) suggest that the number of classes be near $5 \log_{10}(n)$.

Of course, these measures will seldom agree, and should be regarded as guidelines, or rules of thumb. Other considerations will also enter into the choice of the bin width, such as the practical desirability of having the class boundaries fall on values that are "natural" with respect to the data at hand. For example, the January 1987 Ithaca maximum temperature data has IQR = 7°F, and $n = 31$. Class intervals of width 3.4 and 4.5°F would be suggested by Eqs. (3.11a) and (3.11b), respectively. The range of these data is $53 - 9 = 44$°F, so that the numbers of histogram bars suggested by the two equations are about $44/3.4 \approx 13$, and $44/4.5 \approx 10$. The rule of thumb from Panofsky and Brier yields about seven intervals. These three estimates are not too different, and a reasonable choice in this case might be to choose 10 intervals of width 5°F, yielding a histogram looking much like the stem-and-leaf display in Fig. 3.1b.

3.3.6 Cumulative Frequency Distributions

A display related to the histogram is the cumulative frequency distribution, also known as the *empirical cumulative distribution function*. In the older literature, it is also sometimes referred to as the *ogive*. The cumulative frequency distribution is a two-dimensional plot in which the vertical axis shows quantile (i.e., cumulative) probabilities, and the horizontal axis shows the data values themselves. The

Fig. 3.6 Empirical cumulative frequency distribution functions for the January 1987 Ithaca maximum temperature data (a) and precipitation data (b). The S shape exhibited by the temperature data is characteristic of reasonably symmetric data, while the *concave downward* look exhibited by the precipitation data is characteristic of data that are skewed to the right.

plot represents relative frequency estimates for the probability that an arbitrary or randomly chosen datum will not exceed a chosen value on the horizontal axis. Thus, the cumulative frequency distribution is like the integral of a histogram with arbitrarily narrow bin width. Figure 3.6 shows two empirical cumulative distribution functions.

Formally, the empirical cumulative distribution function, $p(x)$, can be expressed as

$$p(x) \approx \Pr\{X \leq x\}. \tag{3.12}$$

The notation on the right side of this equation can be somewhat confusing at first, but is standard in statistical work. The capital letter X represents the generic random variable, or the "arbitrary, randomly chosen" value referred to in the previous paragraph. The lowercase x, on both sides of Eq. (3.12) represents a specific value of the random quantity. In the cumulative frequency distribution, these specific values are plotted on the horizontal axis.

In order to construct a cumulative frequency distribution, it is necessary to estimate $p(x)$ using the ranks, i, of the order statistics, $x_{(i)}$. In the literature of hydrology, these estimates are known as *plotting positions* (e.g., Harter, 1984). One might initially be tempted to use a simple fraction estimate for the function, i.e., $p(x_{(i)}) = i/n$, but on further consideration it can be seen that this is a misleading estimate. In particular, the largest data value in any given sample, $x_{(n)}$, will always be assigned the cumulative probability $n/n = 1.0$. The unwanted implication of

this probability estimate is that, in future batches of comparable data, the probability of observing a value larger than the current $x_{(n)}$ is projected to be zero. In effect, the implication is that the setting of new records is impossible.

A number of alternatives to the estimation of $p(x)$ that do not suffer from this problem are available. Two commonly used choices are

$$p(x_{(i)}) = \frac{i}{n + 1} \tag{3.13a}$$

and

$$p(x_{(i)}) = \frac{i - 1/3}{n + 1/3}. \tag{3.13b}$$

The first of these, Eq. (3.13a), is motivated by the fact that the consecutive order statistics of a random sample are separated, on average, by probability intervals of size $1/(n + 1)$. The second, Eq. (3.13b), is supported by statistical reasoning concerning the median of the possible values of $x_{(i)}$. Further discussion on these and other cumulative probability estimators can be found in Harter (1984) and Hoaglin (1983).

Figure 3.6a shows the cumulative frequency distribution for the January 1987 Ithaca maximum temperature data, using Eq. (3.13b) to estimate the cumulative probabilities. Figure 3.6b shows the Ithaca precipitation data displayed in the same way. For example, the lowest of the 31 temperatures in Fig. 3.6a is $x_{(1)} = 9°$ F, and $p(x_{(1)})$ is plotted at $(1 - .333)/(31 + .333) = 0.0213$. The steepness in the center of the plot reflects the concentration of data values in the center of the distribution, while the flatness at high and low temperatures results from their being more rare. The S-shaped character of this cumulative distribution is indicative of a reasonably symmetrical distribution, with comparable numbers of observations on either side of the median at a given distance from the median. The cumulative distribution function for the precipitation data (Fig. 3.6b) rises quickly on the left, because of the high concentration of data values there, and then rises more slowly in the center and right of the figure because of the relatively fewer large observations. This concave downward look to the cumulative distribution function is indicative of positively skewed data. A plot for a batch of negatively skewed data would show just the reverse characteristics: a very shallow slope in the left and center of the diagram, and rising steeply toward the right, yielding a function that would be concave upward.

3.4 Reexpression

It is possible that the original scale of measurement may obscure important features in a set of data. If so, an analysis can be facilitated, or yield more revealing results if the data are first subjected to a mathematical transformation. Such trans-

formations can also be very useful for helping atmospheric data conform to the assumptions of regression analysis (Chapter 6). In the terminology of exploratory data analysis, such data transformations are known as *reexpression* of the data.

3.4.1 Power Transformations

Often data transformations are undertaken in order to make the distribution of values more nearly symmetrical. The resulting symmetry may allow use of more familiar and traditional statistical techniques. Sometimes a symmetry-producing transformation can make exploratory analyses, such as those described in this chapter, more revealing. These transformations can also aid in comparing batches of data, such as by straightening or smoothing relationships between two variables. Another important use of power transformations is to make the variations or dispersion (i.e., the spread) of one variable less dependent on the value of another variable, in which case the transformation is called *variance stabilizing*.

Undoubtedly the most commonly used symmetry-producing transformations are the two closely related functions

$$T_1(x) = \begin{cases} x^\lambda, & \lambda > 0 \\ \ln(x), & \lambda = 0 \\ -(x^\lambda), & \lambda < 0 \end{cases} \tag{3.14a}$$

and

$$t_2(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases}. \tag{3.14b}$$

These transformations are useful when dealing with unimodal (single-humped) distributions of positive data variables. Each of these functions is a family of transformations indexed by the single parameter λ. The term *power transformation* derives from the fact that the important work of these transformations—changing the shape of the data distribution—is accomplished by the exponentiation, or raising the data values to the power λ. Thus the two sets of transformations in Eqs. (3.14a) and (3.14b) are actually quite comparable, and a particular value of λ produces the same effect on the overall shape of the data in either case. The transformations in Eq. (3.14a) are of a slightly simpler form, and are often employed because it is easier. The transformations in Eq. (3.14b), also known as the *Box–Cox transformations*, are simply shifted and scaled versions of Eq. (3.14a), and are sometimes more useful when comparing among different transformations. Also, Eq. (3.14b) is mathematically "nicer" since the limit of the transformation as $\lambda \to 0$ is actually the function $\ln(x)$.

In both Eqs. (3.14a) and (3.14b), changing the value of the parameter λ yields one of an essentially continuously varying set of smooth transformations. These transformations are sometimes referred to as the *ladder of powers*. A few of

Fig. 3.7 Graphs of the power transformations in Eq. (3.14b) for selected values of the transformation parameter λ. For $\lambda = 1$ the transformation is linear, and produces no change in the shape of the data. For $\lambda < 1$ the transformation reduces all data values, with larger values more strongly affected. The reverse effect is produced by transformations with $\lambda > 1$.

these transformation functions are plotted in Fig. 3.7. The curves in this figure are functions specified by Eq. (3.14b), although the corresponding curves from Eq. (3.14a) have the same shapes. Figure 3.7 makes it clear that use of the logarithmic transformation for $\lambda = 0$ fits neatly into the spectrum of the power transformations. This figure also illustrates another property of the power transformations, which is that they are all increasing functions of the original variable, x. This property is achieved in Eq. (3.14a) by the negative sign in the transformations with $\lambda < 1$. For the transformations in Eq. (3.14b) the sign reversal is achieved by dividing by λ. This strictly increasing property of the power transformations implies that they are order preserving, so that the smallest value in the original data set will correspond to the smallest value in the transformed data set, and likewise for the largest values. In fact, there will be a one-to-one correspondence between all the order statistics of the original and transformed distributions. Thus the median, quartiles, and so on of the original data will become the corresponding quantiles of the transformed data.

Clearly for $\lambda = 1$ the data remain essentially unchanged. For $\lambda > 1$ the data values are increased [possibly except for the subtraction of $1/\lambda$ and division by λ

Fig. 3.8 Effect of a power transformation with $\lambda < 1$ on a batch of data with positive skew (solid curve). Arrows indicate that the transformation moves all the points to the left, with the larger values being moved much more. The resulting distribution (dashed curve) is reasonably symmetrical.

if Eq. (3.14b) is used], with the larger values being increased more than the smaller ones. Therefore power transformations with $\lambda > 1$ will help produce symmetry when applied to negatively skewed data. The reverse is true for $\lambda < 1$, where larger data values are decreased more than smaller values. Power transformations with $\lambda < 1$ are therefore generally applied to data that are originally positively skewed, in order to produce more nearly symmetric data. Figure 3.8 illustrates the mechanics of this process for an originally positively skewed distribution (solid line). Applying a power transformation with $\lambda < 1$ reduces all the data values, but affects the larger values more strongly. An appropriate choice of λ can often produce at least approximate symmetry through this process (dashed distribution). Choosing an excessively small or negative value for λ will produce an overcorrection, and would result in the transformed distribution being negatively skewed.

Initial inspection of an exploratory data plot such as a schematic diagram can indicate quickly the direction and approximate magnitude of the skew in a batch of data. It is thus usually clear whether a power transformation with $\lambda > 1$ or $\lambda < 1$ is appropriate, but a specific value for the exponent will not be so obvious. A number of approaches to choosing an appropriate transformation parameter have been suggested (e.g., Box and Cox, 1964; Emerson and Stoto, 1983; Hinkley, 1977). The simplest of these is the d_λ statistic (Hinkley, 1977),

$$d_\lambda = \frac{|\text{mean} - \text{median}|}{\text{spread}}. \tag{3.15}$$

Here "spread" is some resistant measure of dispersion, such as the IQR or MAD. The Hinkley d_λ is used to decide among power transformations essentially by trial and error, by computing its value for each of a number of different choices for λ. Usually these trial values of λ are spaced at intervals of ½ or ¼. That choice of λ producing the smallest d_λ is then adopted to transform the data. One very easy way to do the computations is with a spreadsheet program on a desk computer.

The basis of the d_λ statistic is that, for symmetrically distributed data, the mean

and median will be very close. Therefore, as successively stronger power transformations (values of λ increasingly far from 1) move the data toward symmetry, the numerator in Eq. (3.15) will move toward zero. As the transformations become too strong, the numerator will begin to increase relative to the spread measure, resulting in the d_λ increasing again.

Example 3.3. Choosing an Appropriate Power Transformation

Table 3.1 shows the 1933–1982 January Ithaca precipitation data from Table A.2 in Appendix A, sorted in ascending order and subjected to the power transformations in Eq. (3.14b), for $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$ and $\lambda = -0.5$. For $\lambda = 1$ this transformation amounts only to subtracting 1 from each data value. Note that even for the negative exponent $\lambda = -0.5$ the ordering of the original

Table 3.1

Transformed Ithaca January Precipitation 1933–1982, from Table A.2.[a]

Year	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = -0.5$	Year	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = -0.5$
1933	−0.56	−0.67	−0.82	−1.02	1948	0.72	0.62	0.54	0.48
1980	−0.48	−0.56	−0.65	−0.77	1960	0.75	0.65	0.56	0.49
1944	−0.46	−0.53	−0.62	−0.72	1964	0.76	0.65	0.57	0.49
1940	−0.28	−0.30	−0.33	−0.36	1974	0.84	0.71	0.61	0.53
1981	−0.13	−0.13	−0.14	−0.14	1962	0.88	0.74	0.63	0.54
1970	0.03	0.03	0.03	0.03	1951	0.98	0.81	0.68	0.58
1971	0.11	0.11	0.10	0.10	1954	1.00	0.83	0.69	0.59
1955	0.12	0.12	0.11	0.11	1936	1.08	0.88	0.73	0.61
1946	0.13	0.13	0.12	0.12	1956	1.13	0.92	0.76	0.63
1967	0.16	0.15	0.15	0.14	1965	1.17	0.95	0.77	0.64
1934	0.18	0.17	0.17	0.16	1949	1.27	1.01	0.82	0.67
1942	0.30	0.28	0.26	0.25	1966	1.38	1.09	0.87	0.70
1963	0.31	0.29	0.27	0.25	1952	1.44	1.12	0.89	0.72
1943	0.35	0.32	0.30	0.28	1947	1.50	1.16	0.92	0.74
1972	0.35	0.32	0.30	0.28	1953	1.53	1.18	0.93	0.74
1957	0.36	0.33	0.31	0.29	1935	1.69	1.28	0.99	0.78
1969	0.36	0.33	0.31	0.29	1945	1.74	1.31	1.01	0.79
1977	0.36	0.33	0.31	0.29	1939	1.82	1.36	1.04	0.81
1968	0.39	0.36	0.33	0.30	1950	1.82	1.36	1.04	0.81
1973	0.44	0.40	0.36	0.33	1959	1.94	1.43	1.08	0.83
1941	0.46	0.42	0.38	0.34	1976	2.00	1.46	1.10	0.85
1982	0.51	0.46	0.41	0.37	1937	2.66	1.83	1.30	0.95
1961	0.69	0.60	0.52	0.46	1979	3.55	2.27	1.52	1.06
1975	0.69	0.60	0.52	0.46	1958	3.90	2.43	1.59	1.10
1938	0.72	0.62	0.54	0.48	1978	5.37	3.05	1.85	1.21

[a] The data have been sorted, with the power transformations in Eq. (3.14b) applied for $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, and $\lambda = 0.5$. For $\lambda = 1$ the transformation subtracts 1 from each data value. Schematic plots of these data are shown in Fig. 3.9.

Fig. 3.9 The effect of power transformations of the form of Eq. (3.14b) on the January total precipitation data for Ithaca, 1933–1982 (Table A.2). The original data ($\lambda = 1$) is strongly skewed to the right, with the largest value being "far out." The square root transformation ($\lambda = 0.5$) improves the symmetry somewhat, whereas the logarithmic transformation ($\lambda = 0$) produces a reasonably symmetrical distribution. When subjected to the more extreme inverse square-root transformation ($\lambda = 0.5$), the data begin to exhibit negative skewness. The logarithmic transformation would be chosen as best by the Hinkley d_λ statistic, Eq. (3.15).

data is preserved in all the transformations. Therefore, in this format it is easy to determine the medians and the quartiles of the original and transformed data.

Figure 3.9 shows schematic plots for the data in Table 3.1. The untransformed data (leftmost plot) are clearly positively skewed, which is usual for distributions of precipitation amounts. All three of the values outside the fences are large amounts, with the largest being "far out." The three other schematic plots show the results of successively stronger power transformations with $\lambda < 1$. The d_λ statistic, computed using IQR as the measure of spread, indicates that the best choice among the three is the logarithmic transformation ($\lambda = 0$). The near symmetry exhibited by the schematic plot for the logarithmically transformed data supports this conclusion. The more extreme inverse square-root transformation ($\lambda = -0.5$) has evidently overcorrected for the positive skewness, as the three smallest amounts are now outside the fences. ¤

3.4.2 Standardized Anomalies

Transformations can also be useful when one is interested in working simultaneously with batches of data that are related, but not strictly comparable. One in-

stance of this situation occurs when the data are subject to seasonal variations. Direct comparison of raw monthly temperatures, for example, will usually show little more than the dominating influence of the seasonal cycle. A record warm January will still be much colder than a record cool July. In situations of this sort, reexpression of the data in terms of *standardized anomalies* can be very helpful.

The standardized anomaly, z, is computed simply by subtracting the sample mean of the batch from which the data are drawn, and dividing by the corresponding sample standard deviation:

$$z = \frac{x - \bar{x}}{s_x} = \frac{x'}{s_x}. \tag{3.16}$$

In the jargon of the atmospheric sciences, an *anomaly* x' is understood to mean the subtraction from a data value of a relevant average, as in the numerator of Eq. (3.16). The term *anomaly* does not connote a data value or event that is abnormal or necessarily even unusual. The standardized anomaly in Eq. (3.16) is produced by dividing the anomaly in the numerator by the corresponding standard deviation. This transformation is sometimes also referred to as a *normalization*. It would also be possible to construct standardized anomalies using resistant measures of location and spread, such as by subtracting the median and dividing by IQR, but this is rarely done. Use of standardized anomalies is motivated by ideas deriving from the bell-shaped Gaussian distribution, which will be made clearer in Chapter 4. However, it is not necessary to assume that a batch of data follows any particular distribution in order to reexpress them in terms of standardized anomalies.

The idea behind the standardized anomaly is to try to remove the influences of location and spread from a batch of data. In the process, the physical units of the original data cancel, so standardized anomalies are always dimensionless quantities. Subtracting the mean produces a series of anomalies, x', located somewhere near zero. Division by the standard deviation puts excursions from the mean in different batches of data on equal footings. Collectively, a batch of data that has been transformed to a set of standardized anomalies will exhibit a mean of zero and a standard deviation of 1.

For example, one often finds that summer temperatures are less variable than winter temperatures. One might find that the standard deviation for average January temperature at some location is around 3°C, whereas the standard deviation for average July temperature at the same location is close to 1°C. An average July temperature 3°C colder than the long-term mean for July would then be quite unusual, corresponding to a standardized anomaly of -3. An average January temperature 3°C warmer than the long-term mean January temperature at the same location would be a fairly ordinary occurrence, corresponding to a standardized anomaly of only $+1$. Another way to look at the standardized anomaly is as

a measure of distance, in standard deviation units, between a data value and its mean.

Example 3.4. Expressing Climatic Data in Terms of Standardized Anomalies

Figure 3.10, taken from Arkin (1989), illustrates the use of standardized anomalies in an operational context. The plotted points are values of an index of the El Niño–Southern Oscillation (ENSO) phenomenon that is used by the Climate Analysis Center of the U.S. National Weather Service (Ropelewski and Jones, 1987). The values of this index in the figure are derived from month-by-month differences in the standardized anomalies of sea-level pressure at two tropical locations: Tahiti, in the central Pacific Ocean; and Darwin, in northern Australia. Thus, in terms of Eq. (3.16) the first step toward computing plotted points is to calculate the difference $\Delta z = z_{\text{Tahiti}} - z_{\text{Darwin}}$ for each month during the 19 years plotted. The standardized anomaly z_{Tahiti} for January 1989, for example, is computed by subtracting the average pressure for all Januaries at Tahiti from the observed monthly pressure for January 1989. This difference is then divided by the standard deviation characterizing the year-to-year variations of January atmospheric pressure at Tahiti.

Actually, the points plotted in Fig. 3.10 are themselves standardized anomalies of this difference of standardized anomalies; thus Eq. (3.16) has been applied twice to the original data. The first of the two standardizations is undertaken to minimize the influences of seasonal changes in the average monthly pressures and the year-to-year variability of those averages. The second standardization, calcu-

Fig. 3.10 Standardized differences between the standardized monthly sea-level pressure anomalies of Tahiti and Darwin, 1970–1988. Solid line indicates the 5-month running mean. [From Arkin (1989), reproduced with permission of the American Meteorological Society.]

lating the standardized anomaly of the difference $z_{\text{Tahiti}} - z_{\text{Darwin}}$, ensures that the resulting index will have unit standard deviation. For reasons that will be made clear in the discussion of the Gaussian distribution in Chapter 4, this attribute aids qualitative judgments about the unusualness of a particular index value.

Physically, during El Niño events the center of tropical Pacific precipitation activity shifts eastward from the western Pacific (near Darwin) to the central Pacific (near Tahiti). This shift is associated with higher-than-average surface pressures at Darwin and lower-than-average surface pressures at Tahiti, which together produce a negative value for the index plotted in Figure 3.10. The exceptionally strong El Niño event of 1982/83 is especially prominent in this figure. A very readable summary of the ENSO phenomenon can be found in Peixoto and Oort (1992). ¤

3.5 Exploratory Techniques for Paired Data

The techniques presented so far in this chapter have pertained mainly to the manipulation and investigation of single batches of data. Some comparisons have been made, such as the side-by-side schematic plots in Fig. 3.5. There the overall distributions of data from Appendix A were compared, but potentially important aspects of the structure of that data were not shown. In particular, the relationships between variables observed on a given day were masked when the data from each batch were separately ranked prior to construction of schematic plots. However, for each observation in one batch there is a corresponding observation from the same date in the other. That is, the observations are *paired*. Elucidating relationships among sets of data pairs often yields important insights.

3.5.1 Scatterplots

The nearly universal format for graphically displaying paired data is the familiar scatterplot, or x–y plot. Geometrically, a *scatterplot* is simply a collection of points in the plane whose two Cartesian coordinates are the values of each member of the data pair. Scatterplots allow easy examination of such features in the data as trends, curvature in the relationship, clustering of one or both variables, changes of spread of one variable as a function of the other, and extraordinary points or outliers.

Figure 3.11 is a scatterplot of the maximum and minimum temperatures for Ithaca during January 1987. It is immediately apparent that very cold maxima are associated with very cold minima, and there is a tendency for the warmer maxima to be associated with the warmer minima. This scatterplot also shows that the central range of maximum temperatures is not strongly associated with minimum

Fig. 3.11 Scatterplot for daily maximum and minimum temperatures during January 1987 at Ithaca, New York. Closed circles represent days with at least 0.01 in. of precipitation (liquid equivalent).

temperature, since maxima near 30° F occur with minima anywhere in the range of −5 to 20° F, or warmer.

Also illustrated in Fig. 3.11 is a useful embellishment on the scatterplot: the use of more than one type of plotting symbol. Here points representing days on which at least 0.01 in. (liquid equivalent) of precipitation was recorded are plotted using the filled circles. As was evident in Example 2.1 concerning conditional probability, precipitation days tend to be associated with warmer minimum temperatures. The scatterplot indicates that the maximum temperatures tend to be warmer as well, but that the effect is not as pronounced.

3.5.2 Pearson ("Ordinary") Correlation

Often an abbreviated, single-valued measure of association between two variables, say, x and y, is needed. In such situations, data analysts almost automatically (and sometimes fairly uncritically) calculate a correlation coefficient. Usually, the term *correlation coefficient* is used to mean the "Pearson product-moment coefficient of linear correlation" between two variables x and y.

One way to view the Pearson correlation is as the ratio of the sample covariance of the two variables to the product of the two standard deviations

$$
\begin{aligned}
r_{xy} &= \frac{\operatorname{cov}(x, y)}{s_x s_y} \\
&= \frac{[1/(n-1)] \sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\{[1/(n-1)] \sum_{i=1}^{n} (x_i - \bar{x})^2\}^{1/2} \{[1/(n-1)] \sum_{i=1}^{n} (y_i - \bar{y})^2\}^{1/2}} \\
&= \frac{\sum_{i=1}^{n} [x_i' y_i']}{[\sum_{i=1}^{n} (x_i')^2]^{1/2} [\sum_{i=1}^{n} (y_i')^2]^{1/2}}, \qquad (3.17)
\end{aligned}
$$

where the primes denote anomalies, or subtraction of mean values, as before. Note that the sample variance is a special case of the covariance [numerator in the first two lines of Eq. (3.17)], with $x = y$. One application of the covariance is in the mathematics used to describe turbulence, where the average product of the anomalies x' and y', called the *eddy covariance*, is used to characterize turbulent transport (e.g., Stull, 1988).

The Pearson product–moment correlation coefficient is neither robust nor resistant. It is not robust because strong but nonlinear relationships between the two variables x and y may not be recognized. It is not resistant since it can be extremely sensitive to one or a few outlying point pairs.

The Pearson correlation has two important properties. First, it is bounded by -1 and 1, i.e., $-1 \leq r_{xy} \leq 1$. If $r_{xy} = -1$, there is a perfect, negative linear association between x and y; that is, the scatterplot of y versus x consists of points all falling along one line, and that line has negative slope. Similarly, if $r_{xy} = 1$, there is a perfect positive linear association. The second important property is that the square of the Pearson correlation, r_{xy}^2, specifies the proportion of the variability of one of the two variables that is linearly accounted for, or described, by the other. Commonly one hears that r_{xy}^2 is the amount of the variance of one variable "explained" by the other, but this expression is imprecise at best and is sometimes misleading. The correlation coefficient provides no explanation at all about the relationship between the variables x and y, at least not in any physical sense. It may be that x physically causes y or vice versa, but often both result physically from some other or many other quantities or processes.

The heart of the Pearson correlation coefficient is the covariance between x and y in the numerator. The denominator is in effect just a scaling constant, and is always positive. Thus, the Pearson correlation is essentially a nondimensionalized covariance. Consider the hypothetical cloud of (x, y) data points in Fig. 3.12, recognizable immediately as exhibiting positive correlation. The two perpendicular lines passing through the two sample means define four quadrants, labeled conventionally using Roman numerals. For each of the n points, the quantity within the summation in the numerator of Eq. (3.17) is calculated. For points in quadrant I, both the x and y values are larger than their respective means ($x' > 0$

Fig. 3.12 Hypothetical cloud of points in two dimensions, illustrating the mechanics of the Pearson correlation coefficient [Eq. (3.17)]. The two sample means divide the plane into four quadrants, numbered I–IV.

and $y' > 0$), so that both factors inside the square brackets will be positive. Therefore points in quadrant I contribute positive terms to the sum in the numerator of Eq. (3.17). Similarly, for points in quadrant III, both x and y are smaller than their respective means ($x' < 0$ and $y' < 0$), and again their product will be positive. Thus points in quadrant III will also contribute positive terms to the sum in the numerator. For points in quadrants II and IV, one of the two variables x and y is above its mean and the other is below. Therefore the product in the numerator of Eq. (3.17) will be negative for points in quadrants II and IV, and these points will contribute negative terms to the sum.

In Fig. 3.12 most of the points are in either quadrants I or III, and therefore most of the terms in the numerator of Eq. (3.17) are positive. Only the two points in quadrants II and IV contribute negative terms, and these are small in absolute value since the x and y values are close to their respective means. The result is a positive sum in the numerator and therefore a positive covariance. The two standard deviations in the denominator of Eq. (3.17) must always be positive, which yields a positive correlation coefficient overall in Fig. 3.12. If most of the points were in quadrants II and IV the point cloud would slope downward rather than upward, and the correlation coefficient would be negative. If the point cloud were more-or-less evenly distributed among the four quadrants the correlation coefficient would be near zero, since the positive and negative terms in the sum in the numerator of Eq. (3.17) would tend to cancel.

Another way of looking at the Pearson correlation coefficient is by moving the scaling constants in the denominator (the standard deviations), inside the summation of the numerator. This operation gives

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} \right] = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}, \quad (3.18)$$

so that another way of looking at the Pearson correlation is as (nearly) the average product of the variables after conversion to standardized anomalies.

From the standpoint of computational economy, the formulas presented so far for the Pearson correlation are awkward. This is true regardless of whether the computation is to be done by hand or by a computer program. In particular, all of them require two passes through the data set before the result is achieved: the first to compute the sample means, and the second to accumulate the terms involving deviations from the data values from their sample means (the anomalies). Passing twice through a data set provides double the opportunity for keying errors when using a hand calculator, and can amount to substantial increases in computer time if working with large data sets. Therefore, it is often useful to know the *computational form* of the Pearson correlation, which allows its computation with only one pass through a data set.

The computational form arises through an easy algebraic manipulation of the summations in the correlation coefficient. Consider the numerator in Eq. (3.17). Multiplying out the terms yields

$$\begin{aligned}
\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})] &= \sum_{i=1}^{n} [x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y}] \\
&= \sum_{i=1}^{n} (x_i y_i) - \bar{y} \sum_{i=1}^{n} (x_i) - \bar{x} \sum_{i=1}^{n} (y_i) + \bar{x}\bar{y} \sum_{i=1}^{n} (1) \\
&= \sum_{i=1}^{n} (x_i y_i) - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} (x_i y_i) - \frac{1}{n} \left[\sum_{i=1}^{n} (x_i) \right] \left[\sum_{i=1}^{n} (y_i) \right]. \qquad (3.19)
\end{aligned}$$

The second line in Eq. (3.19) is arrived at through the realization that the sample means are constant, once the individual data values are determined, and can therefore be moved (factored) outside the summations. In the last term on this line there is nothing left inside the summation except the number 1, and the sum of n of these is simply n. The third step recognizes that the sample size multiplied by the sample mean yields the sum of the data values, which follows directly from the definition of the sample mean [Eq. (1.1)]. The fourth step simply substitutes again the definition of the sample mean, to emphasize that all the quantities necessary for computing the numerator of the Pearson correlation can be known after one pass through the data. These are: the sum of the x values, the sum of the y values, and the sum of their products.

It should be apparent from the similarity in form of the summations in the denominator of the Pearson correlation that analogous formulas can be derived for

them or, equivalently, for the sample standard deviation. The mechanics of the derivation are exactly as followed in Eq. (3.19); the result is

$$s_x = \left[\frac{\Sigma\,(x_i^2) - n\bar{x}^2}{n-1}\right]^{1/2}. \tag{3.20}$$

A similar result is, of course, obtained for y. Mathematically, Eq. (3.20) is exactly equivalent to the formula for the sample standard deviation in Eq. (3.5). Thus Eqs. (3.19) and (3.20) can be substituted into the form of the Pearson correlation given in Eq. (3.17) or (3.18), to yield the computational form for the correlation coefficient

$$r_{xy} = \frac{\Sigma_{i=1}^n x_i y_i - (1/n)(\Sigma_{i=1}^n x_i)(\Sigma_{i=1}^n y_i)}{[\Sigma_{i=1}^n x_i^2 - (1/n)(\Sigma_{i=1}^n x_i)^2]^{1/2}[\Sigma_{i=1}^n y_i^2 - (1/n)(\Sigma_{i=1}^n y_i)^2]^{1/2}}. \tag{3.21}$$

It is important to mention a cautionary note regarding the computational forms just derived. There is a potential problem inherent in their use, which stems from the fact that they are very sensitive to roundoff errors. The problem arises because these formulas involve the difference between two numbers of comparable magnitude. To illustrate, suppose that the two terms on the last line of Eq. (3.19) have each been saved to five significant digits. If the first three of these are the same, their difference will then be known to only two significant digits rather than five. The remedy to this potential problem is to retain as many as possible (preferably all) of the significant digits in each calculation.

Example 3.5. Some Limitations of Linear Correlation

Consider the two artificial data sets in Table 3.2. The data values are few and small enough that the computational form of the Pearson correlation can be used

Table 3.2

Artificial Paired Data Sets for Correlation

Set I		Set II	
x	y	x	y
0	0	2	8
1	3	3	4
2	6	4	9
3	8	5	2
5	11	6	5
7	13	7	6
9	14	8	3
12	15	9	1
16	16	10	7
20	16	20	17

Fig. 3.13 Scatterplots of the two artificial sets of paired data in Table 3.2. Pearson correlation for the data in panel a (set I in Table 3.2) of only 0.88 underrepresents the strength of the relationship, illustrating that this measure of correlation is not robust. Pearson correlation for the data in panel b (set II) is 0.61, reflecting the overwhelming influence of the single outlying point, and illustrating lack of resistance.

without discarding any significant digits. For set I, the Pearson correlation is $r_{xy} = +0.88$, and for set II the Pearson correlation is $r_{xy} = +0.61$. Thus moderately strong linear relationships appear to be indicated for both sets of paired data.

The Pearson correlation is neither robust nor resistant, and these two small data sets have been constructed to illustrate these deficiencies. Figure 3.13 shows scatterplots of the two data sets, with set I in panel a and set II in panel b. For set I the relationship between x and y is actually stronger than indicated by the linear correlation of 0.88. The data points all fall very nearly on a smooth curve, but since that curve is not a straight line, the Pearson coefficient underestimates the strength of the relationship. It is thus not robust to deviations from linearity in a relationship.

Figure 3.13b illustrates that the Pearson correlation coefficient is not resistant to outlying data. Except for the single outlying point, the data in set II exhibit very little structure. If anything, these remaining nine points are weakly negatively correlated. However, the values $x = 20$ and $y = 17$ are so far from their respective sample means that the product of the resulting two large positive differences in the numerator of Eq. (3.17) or (3.18) dominates the entire sum, and erroneously indicates a moderately strong positive relationship among the 10 data pairs overall. ¤

3.5.3 Rank Correlation

A robust and resistant alternative to the Pearson product–moment correlation coefficient is the Spearman *rank correlation coefficient.* The Spearman correlation is simply the Pearson correlation coefficient computed using the ranks of the data. Conceptually, either Eq. (3.17) or Eq. (3.18) is applied, but to the ranks of the data rather than to the data values themselves. For example, consider the first data pair, (2, 8), in set II of Table 3.2. Here $x = 2$ is the smallest of the 10 values of X and

therefore has rank 1. Being the eighth smallest of the 10, $y = 8$ has rank 8. Thus this first data pair would be transformed to (1, 8) before computation of the correlation. Similarly, both x and y values in the outlying pair (20, 17) are the largest of their respective batches of 10, and would be transformed to (10, 10) before computation of the Spearman correlation coefficient.

In practice it is not necessary to use Eq. (3.17), (3.18), or (3.12) to compute the Spearman rank correlation. Rather, the computations are simplified because one always knows in advance what the transformed values will be. Because the "data" are ranks, they consist simply of all the integers from 1 to the sample size n. For example, the average of the ranks of any of the four data batches in Table 3.2 is $(1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10)/10 = 5.5$. Similarly, the standard deviation [Eq. (3.20)] of these first 10 positive integers is about 3.028. Taking advantage of this information, computation of the Spearman rank correlation can be simplified to

$$r_{\text{rank}} = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}, \tag{3.22}$$

where D_i is the difference in ranks between the ith pair of data values. In cases of "ties," where a particular data value appears more than once, all of these equal values are assigned their average rank before computing the D_i values.

Example 3.6. Comparison of Rank Correlations for the Table 3.2 Data

In set I of Table 3.2, there is a monotonic relationship between x and y, so that both batches of data are already arranged in ascending order. Therefore, both members of each of the n pairs has the same rank within its own batch, and the differences D_i are all zero. Actually, the two largest y values are equal, and each would be assigned the rank 9.5. Other than this tie, the sum in the numerator of the second term in Eq. (3.22) is then zero, and the rank correlation is essentially 1. This result better reflects the strength of the relationship between x and y than does the Pearson correlation of 0.88. Thus, while the Pearson correlation coefficient reflects the strength of linear relationships, the Spearman rank correlation reflects the strength of monotonic relationships.

For the data in set II, the X values are presented in ascending order, but the Y values with which they are paired are jumbled. The difference of ranks for the first record is $D_1 = 1 - 8 = -7$. There are only three data pairs in set II for which the ranks match (the fifth, sixth, and the outliers of the tenth pair). The remaining seven pairs will contribute nonzero terms to the sum in Eq. (3.22), yielding $r_{\text{rank}} = 0.018$ for set II. This result reflects much better the very weak overall relationship between X and Y in set II than does the Pearson correlation of 0.61. ¤

3.5.4 Serial Correlation

In Chapter 2 meteorological *persistence*, or the tendency for weather in successive time periods to be similar, was illustrated in terms of conditional probabilities for

33	32	30	29	25	30	53	• • •	17	26	27	30	34		
	33	32	30	29	25	30	53	• • •	17	26	27	30	34	

Fig. 3.14 Construction of a shifted time series of January 1987 Ithaca maximum temperature data. Shifting the data by one day leaves 30 data pairs (enclosed in the box) with which to calculate the lag-1 autocorrelation coefficient.

the two discrete events "precipitation" and "no precipitation." For continuous variables (e.g., temperature) persistence is typically characterized in terms of serial correlation, or temporal autocorrelation. The prefix *auto-* in autocorrelation denotes the correlation of a variable with itself, so that temporal autocorrelation indicates the correlation of a variable with its own future and past values. Sometimes such correlations are referred to as *lagged* correlations. Almost always, autocorrelations are computed as Pearson product–moment correlation coefficients, although there is no reason why lagged ranked correlations cannot be computed as well.

The process of computing autocorrelations can be visualized by imagining two copies of a sequence of observations being written, with one of the series shifted by one unit of time. This shifting is illustrated in Fig. 3.14, using the January 1987 Ithaca maximum temperature data from Table A.1. These data have been rewritten, with the middle part of the month represented by ellipses, on the first line. The same record has been recopied on the second line, but shifted to the right by one day. This results in 30 pairs of temperatures within the box, which are available for the computation of the correlation coefficient.

Autocorrelations are computed by substituting the lagged data pairs into the formula for the Pearson correlation [Eq. (3.17)]. For the lag-1 autocorrelation there are $n - 1$ such pairs. The only real confusion arises because the mean values for the two series will be slightly different. In Fig. 3.14, for example, the mean of the 30 boxed values in the upper series is 29.77°F, whereas that for the boxed values in the lower series is 29.73°F. This difference arises because the upper series does not include the temperature for January 1 but the lower series does not include the temperature for January 31. Denoting the sample mean of the first $n - 1$ values with the subscript "$-$" and that of the last $n - 1$ values with the subscript "+," we can express the lag-1 autocorrelation by

$$r_1 = \frac{\sum_{i=1}^{n-1} [(x_i - \bar{x}_-)(x_{i+1} - \bar{x}_+)]}{[\sum_{i=1}^{n-1} (x_i - \bar{x}_-)^2 \sum_{i=2}^{n} (x_i - \bar{x}_+)^2]^{1/2}}. \tag{3.23}$$

For the January 1987 Ithaca maximum temperature data, for example, $r_1 = 0.52$.

While the lag-1 autocorrelation is the most commonly computed measure of persistence, it is also sometimes of interest to compute autocorrelations at longer lags. Conceptually, this is no more difficult than the procedure for the lag-1 autocorrelation, and computationally the only difference is that the two series are shifted by more than one time unit. Of course, as a time series is shifted increas-

ingly relative to itself there is progressively less overlapping data to work with. Equation (3.23) can be generalized to the lag-k autocorrelation coefficient using

$$r_k = \frac{\sum_{i=1}^{n-k} [(x_i - \bar{x}_-)(x_{i+k} - \bar{x}_+)]}{[\sum_{i=1}^{n-k} (x_i - \bar{x}_-)^2 \sum_{i=k+1}^{n} (x_i - \bar{x}_+)^2]^{1/2}}. \tag{3.24}$$

Here the subscripts "$-$" and "$+$" indicate sample means over the first and last $n - k$ data values, respectively. Equation (3.24) is valid for $0 \leq k < n - 1$, although usually only the lowest few values of k will be of interest. So much data is lost at large lags that correlations for roughly $k > n/2$ or $k > n/3$ are rarely computed.

In situations where a long data record is available it is sometimes acceptable to use an approximation to Eq. (3.24), which simplifies the calculations and allows use of a computational form. In particular, if the data series is sufficiently long, the overall sample mean will be very close to the subset averages of the first and last $n - k$ values. The overall sample standard deviation will be close to the two subset standard deviations for the first and last $n - k$ values as well. Invoking these assumptions leads to the approximation

$$r_k \approx \frac{\sum_{i=1}^{n-k} [(x_i - \bar{x})(x_{i+k} - \bar{x})]}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n-k} (x_i x_{i+k}) - (n - k)\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{3.25}$$

This form for the computation of lag correlations is very commonly used.

3.5.5 Autocorrelation Function

Together, the collection of autocorrelations computed for various lags are called the *autocorrelation function.* Often autocorrelation functions are displayed graphically, with the autocorrelations plotted as a function of lag. Figure 3.15 shows the first seven values of the autocorrelation function for the January 1987

Fig. 3.15 Sample autocorrelation function for the January 1987 Ithaca maximum temperature data. The correlation is 1 for $k = 0$, since the unlagged data are perfectly correlated with themselves. The autocorrelation function decays to essentially zero for $k \geq 5$.

Ithaca maximum temperature data. An autocorrelation function always begins with $r_0 = 1$, since any unshifted series of data will exhibit perfect correlation with itself. It is typical for an autocorrelation function to exhibit a more-or-less gradual decay toward zero as the lag k increases, reflecting the generally weaker statistical relationships between data points further removed from each other in time. It is instructive to relate this observation to the context of weather forecasting. If the autocorrelation function did not decay toward zero after a few days, making reasonably accurate forecasts at that range would be very easy: simply forecasting today's observation (the persistence forecast) or some modification of today's observations would give good results.

The existence of autocorrelation in meteorological and climatological data has important implications regarding the applicability of some standard statistical methods to atmospheric data. In particular, uncritical application of classical methods requiring independence of data within a sample will often give badly misleading results. In some cases it is possible to successfully modify these techniques, by accounting for the temporal correlation using sample autocorrelations. This topic will be discussed in Chapter 5.

3.6 Exploratory Techniques for Higher-Dimensional Data

When exploration, analysis, or comparison of matched data consisting of more than two batches is required, the methods presented so far are not fully adequate. Some exploratory methods available to deal with these more complex situations are outlined in the following sections.

Meteorologists routinely confront this type of problem in operational work, and some unique approaches to multivariate (i.e., concerning simultaneously more than two variables) exploratory data analysis have been developed within the atmospheric sciences. Consider, for example, Fig. 3.16. Individuals trained in meteorology will immediately recognize this multivariate graphical exploratory data-analytic device, although they might not recognize this description of it. The specimen drawn in Fig. 3.16 simultaneously displays the values of seven vari-

Fig. 3.16 A meteorological *station model*, presenting a simultaneous graphical depiction of seven quantities. When plotted on a map, two location variables (latitude and longitude) are added as well, increasing the dimensionality of the depiction to nine quantities.

ables. Some of these are continuous variables (e.g., temperature, dewpoint), and some are discrete quantities (cloud cover, present weather). Two additional variables (latitude and longitude) are added when a collection of these graphics are plotted on a map, bringing the total to nine simultaneously displayed variables.

3.6.1 The Correlation Matrix

A very useful device for simultaneously displaying correlations among more than two batches of matched data is the correlation matrix. For example, the data set in Table A.1 contains six batches of matched data. Correlation coefficients can be computed for each of the 15 distinct pairings of these six variables. In general, for some number I batches of matched data, there are $(I)(I - 1)/2$ distinct pairings of the batches.

The correlation matrix is a two-dimensional array of correlations that systematically organizes the collection of correlation coefficients. It is a square array, with as many rows and columns as there are batches of matched data whose relationships are to be summarized. Each entry in the array, $r_{i,j}$, is indexed by the two subscripts, i and j, that point to the identity of the two variables whose correlation is represented. For example, $r_{2,3}$ would denote the correlation between the second and third variables in a list. The rows and columns in the correlation matrix are numbered correspondingly, so that the individual correlations are arranged as shown in Fig. 3.17.

$$
[\mathbf{R}] = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & \cdots & r_{1,J} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & \cdots & r_{2,J} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & \cdots & r_{3,J} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & \cdots & r_{4,J} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{I,1} & r_{I,2} & r_{I,3} & r_{I,4} & \cdots & r_{I,J} \end{bmatrix}
$$

Row number, i

Column number, j

Fig. 3.17 The layout of a correlation matrix, [**R**]. Correlations $r_{i,j}$ between all possible pairs of $I = J$ variables are arranged so that the first subscript, i, matches the row number, and the second subscript, j, matches the column number.

Table 3.3

Correlation Matrices for the Data in Table A.1[a]

	Ith.,[b] ppt	Ith., max	Ith., min	Can.,[c] ppt	Can., max	Ith., ppt	Ith., max.	Ith., min	Can., ppt	Can., max
Ith., max	−.024					.319				
Ith., min	.287	.718				.597	.761			
Can., ppt	.965	.018	.267			.750	.281	.546		
Can., max	−.039	.957	.762	−.015		.267	.944	.749	.187	
Can., min	.218	.761	.924	.188	.810	.514	.790	.916	.352	.776

[a] Only the lower triangles of the matrices are shown, to omit redundancies and the uninformative diagonal values. The left matrix contains Pearson product–moment correlations, and the right matrix contains Spearman rank correlations.

[b] Ithaca.

[c] Canandaigua.

The correlation matrix was not designed for exploratory data analysis, but rather as a notational shorthand that allows manipulation of the correlations in the framework of linear algebra. (The basics of linear algebra are reviewed in Section 9.2.) As such, parts of the correlation matrix are redundant, and some are simply uninformative. Consider first the diagonal elements of the matrix, i.e., $r_{1,1}$, $r_{2,2}$, $r_{3,3}$, . . . , $r_{I,I}$. These are the correlations of each of the variables with themselves, and are always equal to 1. Second, realize that the correlation matrix is symmetrical. That is, the correlation $r_{i,j}$ between variables i and j is exactly the same number as the correlation $r_{j,i}$ between the same pair of variables, so that the correlation values above and below the diagonal of 1 values are mirror images of each other. Therefore, as noted above, only $(I)(I - 1)/2$ of the I^2 entries in the correlation matrix provide distinct information.

Table 3.3 shows correlation matrices for the data in Table A.1. The matrix on the left contains Pearson product-moment correlation coefficients, and the matrix on the right contains Spearman rank correlation coefficients. As is consistent with usual practice when using correlation matrices for display rather than computational purposes, only the lower triangles of each matrix are actually printed. Omitted are the uninformative diagonal elements and the redundant upper triangular elements. Only the $(6)(5)/2 = 15$ distinct correlation values are presented.

A variety of features in the underlying data can be discerned by studying and comparing these two correlation matrices. First, notice that the six correlations involving only temperature variables have comparable values in both matrices. The strongest Spearman correlations are between like temperature variables at the two locations. Correlations between maximum and minimum temperatures at the same location are moderately large, but weaker. The correlations involving one or both of the precipitation variables differ substantially between the two correlation matrices. There are only a few very large precipitation amounts for each of the

two locations, and these tend to dominate the Pearson correlations, as explained previously. On the basis of this comparison between the correlation matrices, one would therefore suspect that the precipitation data contained some outliers, even without the benefit of knowing the type of data, or of having seen the individual numbers. The rank correlations would be expected to better reflect the degree of association for data pairs involving one or both of the precipitation variables. Subjecting the precipitation variables to an appropriate normalizing transformation would produce no changes in the matrix of Spearman correlations, but would be expected to improve the agreement between the Pearson and Spearman correlations.

3.6.2 The Scatterplot Matrix

The *scatterplot matrix* is a graphical extension of the correlation matrix. The physical arrangement of the correlation coefficients in a correlation matrix is convenient for quick comparisons of relationships between pairs of data batches, but distilling the relationships down to a single number such as a correlation coefficient inevitably hides important details. A scatterplot matrix is an arrangement of individual scatterplots according to the same logic as governs the placement of individual correlation coefficients in a correlation matrix.

Figure 3.18 is a scatterplot matrix for the January 1987 data, with the scatterplots arranged in the same pattern as the correlation matrices in Table 3.3. The complexity of a scatterplot matrix can be bewildering at first, but a large amount of information about the joint behavior of the data is displayed very compactly. For example, quickly evident from a scan of the precipitation rows and columns is the fact that there are just a few large precipitation amounts at each of the two locations. Looking vertically along the column for Ithaca precipitation, or horizontally along the row for Canandaigua precipitation, the eye is drawn to the largest few data values, which appear to line up. Most of the precipitation points correspond to small amounts and therefore hug the opposite axes. Focusing on the plot of Canandaigua versus Ithaca precipitation, it is apparent that the two locations received most of their precipitation for the month on the same few days. Also evident is the association of precipitation with milder minimum temperatures that was seen in a previous look at this same data. The closer relationships between maximum and maximum, or minimum and minimum temperature variables at the two locations—as compared to the maximum versus minimum temperature relationships—can also be seen clearly.

The scatterplot matrix in Fig. 3.18 has been drawn without the diagonal elements in the positions that correspond to the unit correlation of a variable with itself in a correlation matrix. A scatterplot of any variable with itself would be equally dull, consisting only of a straight-line collection of points. However, it is possible to use the diagonal positions in a scatterplot matrix to portray useful univariate information about the variable corresponding to that matrix position.

Fig. 3.18 Scatterplot matrix for the January 1987 data in Table A.1 of Appendix A.

One simple choice would be schematic plots of each variable in the diagonal positions. Another potentially useful choice is the *normal probability plot* (see Chapter 4), which is a graphical comparison of the data with the bell-shaped Gaussian distribution.

The scatterplot matrix can be even more revealing if constructed on an interactive computer system allowing "brushing" of data points in related plots. When brushing, the analyst can select a point or set of points in one plot, and the corresponding points in the same data record then light up or are otherwise differentiated in that plot as well as in all the other plots then visible. For example, when preparing Fig. 3.11 the differentiation of Ithaca temperatures occurring on days with measurable precipitation was achieved by brushing another plot (that plot was not reproduced in Fig. 3.11) involving the Ithaca precipitation values. The solid circles in Fig. 3.11 thus constitute a temperature scatterplot conditional on nonzero precipitation. Brushing can also sometimes reveal surprising relationships in the data by keeping in motion the computer pointing device ("mouse") that does the brushing. The resulting "movie" of brushed points in the other si-

multaneously visible plots essentially enables the analyst to use an additional dimension—time—to differentiate relationships in the data.

3.6.3 Correlation Maps

Correlation matrices such as that in Table 3.3 are understandable and informative, so long as the number of quantities represented (six, in the case of Table 3.3) remains reasonably small. When the number of variables becomes large it may not be possible to easily make sense of the individual values, or even to fit their correlation matrix on a single page. A frequent cause of atmospheric data being excessively numerous for effective display in a correlation matrix is the necessity of working with data from a large number of locations. In this case the geographic arrangement of the locations can be used to organize the correlation information in map form.

Consider, for example, summarization of the correlations among surface pressure at perhaps 200 locations around the world. By the standards of the discipline, this would be only a modestly large set of data. However, this many batches of pressure data would lead to $(200)(199)/2 = 19{,}100$ distinct station pairs, and as many correlation coefficients. A technique that has been used successfully in such situations is construction of a series of *one-point correlation maps*.

Figure 3.19, taken from Bjerknes (1969), is a one-point correlation map for annual surface pressure data. Displayed on this map are contours of Pearson correlations between the pressure data at roughly 200 locations with that at Djakarta, Indonesia. Djakarta is thus the "one point" in this one-point correlation map. Essentially, the quantities being contoured are the values in the row (or column) corresponding to Djakarta in the very large correlation matrix containing all the 19,100 or so correlation values. A complete representation of that large correlation matrix in terms of one-point correlation maps would require as many maps as stations, or in this case about 200. However, not all the maps would be as interesting as Fig. 3.19. The maps for nearby stations (e.g., Darwin, Australia) would look very similar, and many of the interesting features in such maps can be even more compactly displayed.

Clearly Djakarta is located under the "+1.0" on the map, since the pressure data there are perfectly correlated with themselves. Not surprisingly, pressure correlations for locations near Djakarta are quite high, with gradual declines toward zero at locations somewhat further away. This pattern is the spatial analog of the tailing off of the (temporal) autocorrelation function indicated in Fig. 3.15. The surprising feature in Fig. 3.19 is the region in the eastern tropical Pacific, centered on Easter Island, for which the correlations with Djakarta pressure are strongly negative. This negative correlation implies that in years when average pressure at Djakarta (and nearby locations, such as Darwin) are high, pressures in the eastern Pacific are low, and vice versa. This is the expression in the surface pressure data of the ENSO phenomenon, sketched earlier in this chapter, and is an example of

Fig. 3.19 One-point correlation map of annual surface pressures at locations around the globe with those at Djakarta, Indonesia. The strong negative correlation of -0.8 at Easter Island is related to the El Niño–Southern Oscillation phenomenon. [From Bjerknes (1969), reproduced with permission of the American Meteorological Society.]

what has come to be known as a *teleconnection* pattern. In the ENSO *warm phase* the center of tropical Pacific convection moves eastward, producing lower-than-average pressures near Easter Island and higher-than-average pressures at Djakarta. When the precipitation shifts westward during the *cold phase*, pressures are low at Djakarta and high at Easter Island.

Not all geographically distributed correlation data exhibit teleconnection patterns such as the one shown in Fig. 3.19. However, many large-scale fields, especially pressure (or geopotential height) fields, show one or more teleconnection patterns. A device used to simultaneously display these particular aspects of the large underlying correlation matrix is the *teleconnectivity* map. To construct a teleconnectivity map, the row for each station or gridpoint in the correlation matrix is searched for the largest negative value. The teleconnectivity value for location i, T_i, is the absolute value of that most negative correlation:

$$T_i = |\min_j r_{i,j}|. \tag{3.26}$$

Here the minimization over j (the column index for $[\mathbf{R}]$) implies that all correlations $r_{i,j}$ in the ith row of $[\mathbf{R}]$ are searched for the smallest (most negative) value. For example, in Fig. 3.19 the largest negative correlation with Djakarta pressures is with Easter Island, -0.80. The teleconnectivity for Djakarta surface pressure would therefore be 0.80, and this value is plotted on a teleconnectivity map at the location of Djakarta. To construct the full teleconnectivity map for surface pressure, the other 199 or so rows of the correlation matrix, each corresponding to another station, would be examined for the largest negative correlation (or, if none were negative, then the smallest positive one), and its absolute value would be plotted at the map position of that station.

Figure 3.20, from Wallace and Blackmon (1983), shows the teleconnectivity map for Northern Hemisphere winter 500-millibar (mb) heights. The density of the shading indicates the magnitude of the individual gridpoint teleconnectivity values. The locations of local maxima of teleconnectivity are indicated by the positions of the numbers, expressed as $T_i \times 100$. The arrows in Fig. 3.20 point from the teleconnection "centers" to the location with which each maximum negative correlation is exhibited. The unshaded regions include gridpoints for which the teleconnectivity is relatively low. The one-point correlation maps for locations in these unshaded regions would tend to show gradual declines toward zero at increasing distances, analogously to the time correlations shown in Fig. 3.15, but without declining much further to large negative values.

It has become apparent that a fairly large number of these teleconnection patterns exist in the atmosphere, and the many double-headed arrows indicate that these group naturally into patterns. Especially impressive is the four-center pattern arcing from the central Pacific to the southeastern United States, known as the *Pacific–North America* (PNA) pattern. Notice, however, that these patterns emerged here from a statistical, exploratory analysis of a large mass of atmo-

Fig. 3.20 Teleconnectivity, or strongest negative correlation from each of many one-point correlation maps plotted at the base gridpoint, for winter 500-mb heights. [From Wallace and Blackmon (1983).]

spheric data. This type of work actually had its roots in the early part of the century (see Brown and Katz, 1991), and is a good example of exploratory data analysis in the atmospheric sciences turning up interesting patterns in very large data sets. This is especially so because the exploratory work has served to stimulate research aimed at understanding the underlying physical bases of the patterns, and has lead to tools used operationally in long-range forecasting (Epstein, 1988; Wagner, 1989).

Exercises

3.1. Compare the median, trimean, and the mean of the precipitation data in Table A.3.

3.2. Compute the MAD, the IQR, and the standard deviation of the pressure data in Table A.3.

3.3. Draw a stem-and-leaf display for the temperature data in Table A.3.

3.4. Compute the Yule–Kendall Index and the skewness coefficient using the temperature data in Table A.3.

3.5. Draw the empirical cumulative frequency distribution for the pressure data in Table A.3. Compare it with a histogram of the same data.

3.6. Compare the boxplot and the schematic plot representing the precipitation data in Table A.3.

3.7. Use Hinkley's d_λ to find an appropriate power transformation for the precipitation data in Table A.2 using Eq. (3.14a), rather than Eq. (3.14b) as was done in Example 3.3. Use IQR in the denominator of Eq. (3.15).

3.8. Construct side-by-side schematic plots for the candidate, and final, transformed distributions derived in Exercise 3.7. Compare the result to Fig. 3.9.

3.9. Express the June 1951 temperature in Table A.3 as a standardized anomaly.

3.10. Plot the autocorrelation function up to lag 3, for the Ithaca minimum temperature data in Table A.1.

3.11. Construct a scatterplot of the temperature and pressure data in Table A.3.

3.12. Construct correlation matrices for the data in Table A.3 using
(a) The Pearson correlation
(b) The Spearman rank correlation

Chapter 4 Theoretical Probability Distributions

4.1 Background

In Chapter 3 methods for exploring and displaying variations in data sets were presented. These methods had at their heart the expression of how, empirically, a particular set of data are distributed through their range. The methods presented in this chapter involve imposition of particular mathematical forms to represent the distributions of underlying data. These mathematical forms amount to idealizations of real data, and are accordingly sometimes referred to as "theoretical" distributions.

It is worth taking a moment to understand why one would voluntarily commit the violence of shoehorning real data into an abstract mold. The question is worth considering because theoretical distributions *are* abstractions. They will represent real data only approximately, although in many cases the approximation can be very good, indeed. Basically, there are three ways in which employing theoretical probability distributions may be useful:

- Compactness: Particularly when dealing with large data sets, repeatedly manipulating the raw data can be cumbersome, or even severely limiting. A well-fitting theoretical distribution reduces the number of quantities required for characterizing properties of the data from the full n order statistics $[x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}]$ to a few distribution parameters.
- Smoothing and Interpolation: Real data are subject to sampling variations that lead to gaps or rough spots in the empirical distributions. For example, in Figs. 3.1 and 3.6a, there are no maximum temperature values between 10 and 16°F, although certainly maximum temperatures in this range can and do occur during January at Ithaca. Imposition of a theoretical distribution on these data would represent the possibility of these temperatures occurring, as well as allowing estimation of their probabilities of occurrence.
- Extrapolation: Estimating probabilities for events outside the range of a particular data set requires assumptions about as-yet unobserved behavior. Again referring to Fig. 3.6a, the empirical cumulative probability associated with the coldest temperature, 9°F, was plotted as 0.0213. Using Eq. (3.13b), the probability of a maximum temperature this cold or colder would be estimated as

0.0213, but nothing can be said quantitatively about the probability of January maximum temperatures colder than 5 or 0° F without the imposition of a probability model (i.e., a theoretical distribution) for these data.

While the distinction has been drawn between empirical and theoretical data representations, it should be emphasized that use of theoretical probability distributions does not constitute a method that is independent of empirical considerations. In particular, before embarking on the representation of data using theoretical functions one must decide among candidate distribution forms, fit parameters of the chosen distribution, and check that the resulting function does, indeed, provide a reasonable fit. All three of these steps require use of real data.

4.1.1 What Is a Theoretical Distribution?

A theoretical distribution is an abstract mathematical form, or characteristic shape. Some of these mathematical forms arise naturally as a consequence of certain kinds of data-generating processes, and when applicable, these are especially plausible candidates for concisely representing variations in a set of data. Even when there is no strong natural basis behind the choice of a particular theoretical distribution, it may be found empirically that the distribution represents a set of data very well.

The specific nature of a theoretical distribution is determined by particular values for entities called *parameters* of that distribution. Theoretical distributions are also called *parametric distributions*, because their specific attributes depend on the numerical values of their parameters. For example, the Gaussian (or "normal") distribution has as its characteristic shape the familiar symmetrical "bell." However, merely asserting that a particular batch of data, say, average September temperatures at a location of interest, is well represented by the Gaussian distribution is not very informative about the nature of the data, without specifying *which* Gaussian distribution represents the data. There are in fact infinitely many particular examples of the Gaussian distribution, corresponding to all possible values of the two distribution parameters μ and σ. But knowing, for example, that the monthly temperature for September is well represented by the Gaussian distribution with $\mu = 60°$ F and $\sigma = 2.5°$ F conveys a large amount of information about the nature and magnitudes of the variations of September temperature at that location.

4.1.2 Parameters versus Statistics

There is a potential for confusion between the distribution parameters, and sample statistics. Distribution parameters are abstract characteristics of particular theoretical distributions. They succinctly represent underlying "population" properties. By contrast, a *statistic* is any quantity computed from a sample of data. The confusion between them arises because, for some common theoretical distributions, certain sample statistics are good estimators for the distribution parameters.

For example, the sample standard deviation, s [Eq. (3.5)], can be confused with the parameter σ of the Gaussian distribution because the two are equated when finding a particular Gaussian distribution to best match a data sample. Distribution parameters are found (fitted) using sample statistics. However, it is not always the case that the fitting process is as simple as that for the Gaussian distribution, where the sample mean is equated to the parameter μ and the sample standard deviation is equated to the parameter σ. Usually, the notation for sample statistics involves Roman (i.e., ordinary) letters, and that for parameters involves Greek letters.

4.1.3 Discrete versus Continuous Distributions

There are two distinct types of theoretical distributions, corresponding to different types of data, or random variables. Discrete distributions describe random quantities (i.e., the data of interest) that can take on only particular values. That is, the allowable values are finite, or at least countably infinite. For example, a discrete random variable might take on only the values 0 or 1; or any of the nonnegative integers; or one of the colors red, yellow, or blue. A continuous random variable typically can take on any value within a specified range of the real numbers. For example, a continuous random variable might be defined on the real numbers between 0 and 1; or the nonnegative real numbers; or, for some distributions, any real number.

Strictly speaking, using a continuous distribution to represent observable data implies that the underlying observations are known to an arbitrarily large number of significant figures. Of course, this is never true, but it is convenient and not too inaccurate to represent as continuous those variables that are continuous conceptually but reported discretely. Temperature and precipitation are two obvious examples that really range over some portion of the real number line, but which are reported to discrete multiples of 1°F and 0.01 in. in the United States. Little is lost when treating these discrete observations as samples from continuous distributions.

4.2 Discrete Distributions

There exist a large number of theoretical probability distributions for discrete random variables. Many of these are listed in the encyclopedic volume by Johnson and Kotz (1969), together with many results concerning their properties. However, only three discrete probability distributions have been used to any appreciable degree in the atmospheric sciences: the binomial distribution, the geometric distribution, and the Poisson distribution.

4.2.1 Binomial Distribution

The binomial distribution is probably the simplest possible theoretical distribution, and is therefore often employed in textbooks to illustrate the use and prop-

erties of theoretical distributions more generally. This distribution pertains to outcomes of situations where, on some number of occasions (sometimes called *trials*), one or the other of two MECE events will occur. Classically the two events have been called "success" and "failure," but these are arbitrary labels. More generally, one of the events (say the "success") is assigned the number 1, and the other (the "failure") is assigned the number zero.

The random variable of interest, X, is the number of event occurrences (given by the sum of 1 and 0 values) in some number of trials. The number of trials, N, can be any positive integer, and the variable X can take on any of the nonnegative integer values from 0 (if the event of interest does not occur at all in the N trials) to N (if the event occurs on each occasion). The binomial distribution can be used to calculate probabilities for each of these $N + 1$ possible values of X if two conditions are met: (1) the probability of the event occurring does not change from trial to trial and (2) the outcomes on each of the N trials are mutually independent. These conditions are rarely strictly met, but real situations can be close enough to this ideal that the binomial distribution provides sufficiently accurate representations.

One implication of the first restriction, relating to constant occurrence probability, is that events whose probabilities exhibit regular cycles must be treated carefully. For example, the event of interest might be a thunderstorm or a dangerous lightning occurrence, at a location where there is a diurnal or annual variation in the probability of the event. In cases like these, subperiods (e.g., hours or months, respectively) with approximately constant occurrence probabilities would usually be analyzed separately.

The second necessary condition for applicability of the binomial distribution, relating to event independence, is usually more troublesome for atmospheric data. For example, the binomial distribution seldom would be directly applicable to daily precipitation occurrence or nonoccurrence. As illustrated by Example 2.2, there is often substantial day-to-day dependence between such events. For situations like this the binomial distribution can be generalized to a theoretical stochastic process called a *Markov chain*, discussed in Chapter 8, on time-series methods. On the other hand, the year-to-year statistical dependence in the atmosphere is usually weak enough that occurrences or nonoccurrences of an event in consecutive annual periods can be considered to be effectively independent (12-month "climate" forecasts would be much easier if they were not!). An example of this kind will be presented below.

The usual first illustration of the binomial distribution is in relation to coin flipping. If the coin is fair, the probability of either "heads" or "tails" is 0.5, and does not change from one coin-flipping occasion (or, equivalently from one coin) to the next. If $N > 1$ coins are flipped simultaneously, the outcome on one of the coins does not affect the other outcomes. The coin-flipping situation thus satisfies all the requirements for description by the binomial distribution: dichotomous, independent events with constant probability.

Consider a game where $N = 3$ fair coins are flipped simultaneously, and one is

interested in the number, X, of heads that result. The possible values of X are 0, 1, 2, and 3. These four values are a MECE partition of the sample space for X, and their probabilities must therefore sum to 1. In this simple example, you may not need to think explicitly about the binomial distribution to realize that the probabilities for these four events are 1/8, 3/8, 3/8, and 1/8, respectively.

In the general case, probabilities for each of the $N + 1$ values of X are given by the probability distribution function for the binomial distribution,

$$\Pr\{X = x\} = \binom{N}{x} p^x (1 - p)^{N-x}. \tag{4.1}$$

Here, consistent with the usage in Eq. (3.12), the capitalized X indicates the random variable whose precise value is unknown, or has yet to be observed. The lowercase x denotes a specific, particular value that the random variable can take on. The binomial distribution has two parameters, N and p. The parameter p is the probability of occurrence of the event of interest (the "success") on each one of the N independent trials. For a given pair of the parameters N and p, Eq. (4.1) is a function associating a probability with each of the discrete values $x = 0, 1, 2, \ldots, N$, such that $\Sigma_x \Pr\{X = x\} = 1$. Note that the binomial distribution is unusual in that both of its parameters are conventionally represented by Roman letters.

The right-hand side of Eq. (4.1) consists of two parts: a combinatorial part and a probability part. The combinatorial part specifies the number of distinct ways of realizing x "success" outcomes from a collection of N trials. It is pronounced "N choose x," and is computed according to

$$\binom{N}{x} = \frac{N!}{x!(N - x)!}. \tag{4.2}$$

By convention, $0! = 1$. For example, when tossing $N = 3$ coins, there is only one way that $x = 3$ heads can be achieved: All three coins must come up heads. Using Eq. (4.2), "three choose three" is given by $3!/(3!\,0!) = (1 \cdot 2 \cdot 3)/(1 \cdot 2 \cdot 3 \cdot 1) = 1$. There are three ways in which $x = 1$ can be achieved: Either the first, the second, or the third coin can come up heads, with the remaining two coins coming up tails. Using Eq. (4.2) one obtains $3!/(1!\,2!) = (1 \cdot 2 \cdot 3)/(1 \cdot 1 \cdot 2) = 3$.

The probability part of Eq. (4.1) follows from the multiplicative law of probability for independent events, Eq. (2.11). The probability of a particular arrangement of exactly x event occurrences and $N - x$ nonoccurrences is simply p multiplied by itself x times, and then multiplied by $1 - p$ (the probability of nonoccurrence) $N - x$ times.

Example 4.1. Binomial Distribution and the Freezing of Cayuga Lake, I

Consider the data in Table 4.1, which lists years during which Cayuga Lake, in central New York State, was observed to have frozen. Cayuga Lake is rather deep,

Table 4.1

Years in Which Cayuga Lake Has Frozen, as of 1994

1796	1875	1912	1961
1816	1884	1934	1979
1856	1904		

and will freeze only after a long period of exceptionally cold weather. In any given winter, the lake surface either freezes or does not. Whether the lake freezes in a given winter is essentially independent of whether it froze in recent years. Unless there has been appreciable climate change in the region over the past 200 years, the probability that the lake will freeze in a given year is effectively constant. Therefore, one expects the binomial distribution to provide a good statistical description of the freezing of this lake.

In order to use the binomial distribution as a representation of the statistical properties of the lake-freezing data, we need to "fit" the distribution to the data. Fitting the distribution simply means finding particular values for the distribution parameters, p and N in this case, for which Eq. (4.1) will behave like the data in Table 4.1. The binomial distribution is somewhat unique in that the parameter N depends on the question we want to ask, rather than on the data per se. If we want to compute the probability of the lake freezing next winter, or in any single winter in the future, then $N = 1$. [The special case of Eq. (4.1) with $N = 1$ is called the *Bernoulli distribution.*] If we want to compute probabilities for the lake freezing, say, at least once during some decade in the future, then $N = 10$.

The binomial parameter, p, in this application is the probability that the lake freezes in any given year. It is natural to estimate this probability using the relative frequency of the freezing events in the data. This is a straightforward task here, except for the small complication of not knowing exactly when the climatic record starts. The written record clearly starts no later than 1796, but probably began some years before that. Suppose for simplicity that the data in Table 4.1 represent a 200-year record. The 10 observed freezing events then lead to the relative frequency estimate for the binomial p of $10/200 = 0.05$.

We are now in a position to use Eq. (4.1) to estimate probabilities of a variety of events relating to the freezing of this lake. The simplest kinds of events to work with relate to the lake freezing exactly a specified number of times, x, in a specified number of years, N. For example, the probability of the lake freezing exactly once in 10 years is

$$\begin{aligned}\Pr\{X = 1\} &= \binom{10}{1} 0.05^1 (1 - 0.05)^{10-1} \\ &= \frac{10!}{1!\,9!}(0.05)(0.95^9) = 0.32. \qquad ¤ \qquad (4.3)\end{aligned}$$

Example 4.2. Binomial Distribution and the Freezing of Cayuga Lake, II

A somewhat harder class of events to deal with is exemplified by the problem of calculating the probability that the lake freezes at least once in 10 years. It is clear from Eq. (4.3) that this probability will be no smaller than 0.32, since the probability for the compound event will be given by the sum of the probabilities $\Pr\{X = 1\} + \Pr\{X = 2\} + \cdots + \Pr\{X = 10\}$. This follows from Eq. (2.5), and the fact that these events are mutually exclusive: The lake cannot freeze both exactly once and exactly twice in the same decade.

The "brute force" approach to this problem is to calculate all 10 probabilities in the sum, and then add them up. This is rather tedious, however, and quite a bit of work can be saved by giving the problem a bit more thought. Consider that the sample space here is comprised of eleven MECE events: that the lake freezes exactly 0, 1, 2, . . . , or 10 times in a decade. Since the probabilities for these 11 events must sum to 1, it is much easier to proceed using

$$\begin{aligned}\Pr\{X \geq 1\} &= 1 - \Pr\{X = 0\} \\ &= 1 - \frac{10!}{0!\ 10!}(0.05)^0(0.95)^{10} = 0.40. \qquad ¤ \qquad (4.4)\end{aligned}$$

It is worth noting that the binomial distribution can be applied to situations that are not intrinsically binary, through a suitable redefinition of events. For example, temperature is not intrinsically binary and, indeed, is not even intrinsically discrete. However, for some applications it is of interest to consider the probability of frost, i.e., $\Pr\{T \leq 32°\text{F}\}$. Together with the probability of the complementary event, $\Pr\{T > 32°\text{F}\}$, the situation is one concerning dichotomous events, and could therefore be a candidate for representation using the binomial distribution.

4.2.2 Geometric Distribution

The geometric distribution is related to the binomial distribution, describing a different aspect of the same conceptual situation. Both distributions pertain to a collection of independent trials in which one or the other of a pair of dichotomous events occurs. The trials are independent in the sense that the probability of the "success" occurring, p, is unchanged from trial to trial regardless of the outcomes of previous trials. For the geometric distribution to be applicable, the collection of trials must occur in a sequence.

Although the binomial distribution pertains to probabilities that the particular number of successes will be realized in a fixed number of trials, the geometric distribution specifies probabilities for the number of trials that will be required to observe the next "success." For the geometric distribution, this number of trials is the random variable X, and the probabilities corresponding to its possible values are given by the geometric probability distribution function

$$\Pr\{X = x\} = p(1 - p)^{x-1}, \qquad x = 1, 2, \cdots. \tag{4.5}$$

Here X can take on any positive integer value, since at least one trial will be required in order to observe a "success," and it is possible (although vanishingly probable) that one would have to wait indefinitely for this outcome. Equation (4.6) can be viewed as an application of the multiplicative law of probability for independent events, as it multiplies the probability for a "success" by the probability of a sequence of $x - 1$ consecutive "failures."

Often the geometric distribution is applied to trials that occur consecutively through time, so it is sometimes called the "waiting" distribution. The distribution has been used to describe lengths of weather regimes, or "spells." For example, Waymire and Gupta (1981) show how the geometric distribution can be applied to describe sequences of dry time periods, where one is waiting for a wet event; and sequences of wet periods, during which one is waiting for a dry event.

4.2.3 Poisson Distribution

The Poisson distribution describes the numbers of discrete events occurring in a series, or a sequence, exhibiting a particular kind of independence. Usually the sequence is understood to be in time, such as the occurrence of Atlantic Ocean hurricanes during a particular hurricane season, or perhaps occurrences of droughts or cold spells. However, it is also possible to apply the Poisson distribution to counts of events occurring in one or more spatial dimensions, such as the number of gasoline stations along a particular stretch of highway, or the distribution of hailstones in a small area.

The individual events being counted are independent in the sense that their probability of occurrence in a given interval depends only on the size of the interval, usually the length of the time interval over which they will be counted. The numbers of occurrences do not depend on where in time the interval is located or how often events have occurred in other nonoverlapping intervals. Thus Poisson events occur randomly, but at a constant average rate. A sequence of such events is sometimes said to have been generated by a *Poisson process*. As was the case for the binomial distribution, strict adherence to this independence condition is often difficult to demonstrate in atmospheric data, but the Poisson distribution can still yield a useful representation if the degree of dependence is not too strong. Ideally, Poisson events should be rare enough that the probability of more than one occurring simultaneously is very small. Another way of motivating the Poisson distribution is as the limiting case of the binomial distribution, with p approaching zero and N approaching infinity.

The Poisson distribution has a single parameter, μ, which specifies the average occurrence rate. The Poisson parameter is sometimes called the *intensity*, and has physical dimensions of occurrences per unit time. The probability distribution function for the Poisson distribution is

$$\Pr\{X = x\} = \frac{\mu^x e^{-\mu}}{x!}, \tag{4.6}$$

which associates probabilities with all possible numbers of occurrences, X—from zero to infinitely many. Here $e = 2.718 \cdots$ is the base of the natural logarithms. The sample space for Poisson events therefore contains (countably) infinitely many elements. Clearly the summation of Eq. (4.6) for x running from zero to infinity must be convergent, and equal to 1. The probabilities associated with very large numbers of counts are vanishingly small, since the denominator in Eq. (4.6) is $x!$.

To use the Poisson distribution it must be fit to a sample of data. Fitting the distribution means finding the specific value for the parameter μ that makes Eq. (4.6) behave similarly to a set data at hand. For the Poisson distribution, a good way to estimate the parameter μ is by the *method of moments*. To use the method of moments, one mathematically equates the sample moments and the distribution (or "population") moments. Recall that the first moment is simply the mean. Fitting the Poisson distribution is thus especially easy, since its one parameter is the mean number of occurrences per unit time. Thus the Poisson parameter μ is simply estimated as being equal to the sample number of occurrences per unit time.

Example 4.3. Poisson Distribution and Annual Tornado Counts in New York State

Consider the Poisson distribution in relation to the annual tornado counts in New York State for 1959–1988, listed in Table 4.2. During the 30 years covered by these data, 138 tornados were reported in New York State. The average, or mean, rate of tornado occurrence is simply 138/30 = 4.6 tornados/year, so this is

Table 4.2

Numbers of Tornados Reported Annually in New York State, 1959–1988

1959	3	1969	7	1979	3
1960	4	1970	4	1980	4
1961	5	1971	5	1981	3
1962	1	1972	6	1982	3
1963	3	1973	6	1983	8
1964	1	1974	6	1984	6
1965	5	1975	3	1985	7
1966	1	1976	7	1986	9
1967	2	1977	5	1987	6
1968	2	1978	8	1988	5

Fig. 4.1 Histogram of number of tornados reported annually in New York State for 1959–1988 (dashed), and fitted Poisson distribution with $\mu = 4.6$ tornados per year (solid).

the method-of-moments estimate of the Poisson intensity for this data. Having fit the distribution by estimating a value for this parameter, it can now be used to compute probabilities that particular numbers of tornados will be reported in New York annually. The first 13 of these probabilities (pertaining to 0 through 12 tornados per year) are plotted in the form of a histogram in Fig. 4.1, together with a histogram of the actual data.

The Poisson distribution allocates probability smoothly (given the constraint that the data are discrete) among the possible outcomes, with the most probable numbers of tornados being near the mean rate of 4.6. The distribution of the data shown by the dashed histogram resembles that of the fitted Poisson distribution, but is much more irregular, due at least in part to sampling variations. For example, there does not seem to be a physically based reason why four tornados per year should be substantially less likely than three tornados per year, or why two tornados per year should be less likely than only one tornado per year. Fitting the theoretical distribution to these data provides a sensible way to smooth out these variations, which is desirable if the irregular variations in the data histogram are not physically meaningful. Similarly, using the Poisson distribution to summarize the data allows one to estimate probabilities for zero tornados per year, or for greater than nine tornados per year, even though these numbers of tornados do not happen to have been reported during 1959–1988. Since these numbers of tornados do not appear in the data, one could not estimate probabilities for them using, for example, the empirical cumulative distribution function [Eq. (3.13)]. ¤

4.3 Statistical Expectations

The *expected value* of a random variable or function of a random variable is simply the probability-weighted average of that variable or function. This weighted average is called the "expected" value, even though one does not necessarily expect this outcome to occur in the informal sense of an "expected" event being likely. Paradoxically, it can happen that the statistical expected value is an impossible outcome. Statistical expectations are closely tied to probability distributions, since the distributions will provide the weights in the weighted average. The ability to work easily with statistical expectations can be a strong motivation for choosing to represent data using theoretical distributions.

4.3.1 Expected Value of a Random Variable

It is easiest to see expectations as probability-weighted averages in the context of a discrete probability distribution such as the binomial. Conventionally, the expectation operator is denoted E[ ]:

$$E[X] = \sum_x x \Pr\{X = x\}. \tag{4.7}$$

The summation is taken over all allowable values of X. For example, the expected value of X when X follows the binomial distribution is

$$E[X] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1 - p)^{N-x}. \tag{4.8}$$

Here the allowable values of X are the nonnegative integers up to and including N, and each term in the summation consists of the specific value of the variable, x, times the probability of its occurrence from Eq. (4.1).

The expectation $E[X]$ has a special significance, since it is the mean of the distribution of X. Distribution (or "population") means are commonly denoted using the symbol μ. It is possible to simplify analytically Eq. (4.8) to obtain, for the binomial distribution, the result $E[X] = Np$. Thus the mean of any binomial distribution is given by the product $\mu = Np$. Expressions for the means of other distributions are, of course, different: The comparable result for the geometric distribution [Eq. (4.5)] is $\mu = E[X] = 1/p$, and for the Poisson distribution [Eq. (4.6)] it is $E[X] = \mu$. The New York tornado data in Table 4.2 constitute an example of the expected value, $E[X] = 4.6$ tornados, being impossible to realize in any single year.

4.3.2 Expected Value of a Function of a Random Variable

It can be very useful to compute expectations, or probability-weighted averages, of functions of random variables, $E[g(x)]$. Expectations of function of random

variables have the following properties:

$$E[c] = c \tag{4.9a}$$

$$E[cg_1(x)] = cE[g_1(x)] \tag{4.9b}$$

$$E\left[\sum_{j=1}^{J} g_j(x)\right] = \sum_{j=1}^{J} E[g_i(x)], \tag{4.9c}$$

where c is any constant and $g_j(x)$ is any function of x. Since the constant c does not depend on x, $E[c] = \Sigma_x\, c \Pr\{X = x\} = c\, \Sigma_x \Pr\{X = x\} = c \cdot 1 = c$. Equations (4.9a) and (4.9b) reflect the fact that constants can be factored out when computing expectations. Equation (4.9c) expresses the important property that the expectation of a sum is equal to the sum of the separate expected values.

Use of the properties expressed in Eqs. (4.9) can be illustrated with the expectation of the function $g(x) = (x - \mu)^2$. The expected value of this function is called the *variance*. Substituting into Eq. (4.7), multiplying out terms, and applying the properties in Eqs. (4.9) yields

$$\begin{aligned}
\mathrm{Var}[X] = E[(X - \mu)^2] &= \sum_x (x - \mu)^2 \Pr\{X = x\} \\
&= \sum_x (x^2 - 2\mu x + \mu^2) \Pr\{x = x\} \\
&= \sum_x x^2 \Pr\{x = x\} - 2\mu \sum_x x \Pr\{x = x\} \\
&\quad + \mu^2 \sum_x \Pr\{x = x\} \\
&= E[X^2] - 2\mu\, E[X] + \mu^2 \cdot 1 \\
&= E[X^2] - \mu^2. \qquad (4.10)
\end{aligned}$$

Notice the similarity of the first right-hand side in Eq. (4.10) to the sample variance, given by the square of Eq. (3.5). Similarly, the final equality in Eq. (4.10) is analogous to the computational form for the sample variance, given by the square of Eq. (3.20). For the binomial distribution, $\mathrm{Var}[X] = Np(1 - p)$. For the geometric distribution, $\mathrm{Var}[X] = (1 - p)/p^2$. For the Poisson distribution, $\mathrm{Var}[X] = \mu$.

Example 4.4. Expected Value of a Function of a Binomial Random Variable

Table 4.3 illustrates the computation of statistical expectations for the binomial distribution with $N = 3$ and $p = 0.5$. These parameters correspond to the physical situation of simultaneously flipping three coins, and counting X = the number of heads coming up. The first column shows the possible outcomes of X, and the second column shows the probabilities for each outcome, computed according to Eq. (4.1).

The third column in Table 4.3 shows the individual terms in the probability-

Table 4.3

Binomial probabilities for $N = 3$ and $p = .5$ and construction of expectations $E[X]$ and $E[X^2]$ as probability-weighted averages

X	$\Pr(X = x)$	$x \cdot \Pr(X = x)$	$x^2 \cdot \Pr(X = x)$
0	0.125	0.000	0.000
1	0.375	0.375	0.375
2	0.375	0.750	1.500
3	0.125	0.375	1.125
		$E[X] = 1.500$	$E[X^2] = 3.000$

weighted average $E[X] = \Sigma_x[x \Pr(X = x)]$. Adding these four values yields $E[X] = 1.5$, as would be obtained by multiplying the two distribution parameters $\mu = Np$.

The fourth column in Table 4.3 similarly shows the construction of the expectation $E[X^2] = 3.0$. One might imagine this expectation in the context of a hypothetical game, in which the player receives $\$X^2$: nothing if zero heads come up, \$1 if one head comes up, \$4 if two heads come up, and \$9 if three heads come up. Over the course of many rounds of this game, the long-term average payout would be $E[X^2] = \$3.00$. An individual willing to pay more than \$3 to play this game would be either foolish, or inclined toward taking risks.

Notice that the final equality in Eq. (4.10) can be verified for this particular binomial distribution using Table 4.3. Here $E[X^2] - \mu^2 = 3.0 - (1.5)^2 = 0.75$, agreeing with $\mathrm{Var}[X] = Np(1 - p) = 3(0.5)(1 - 0.5) = 0.75$. ¤

4.4 Continuous Distributions

Most atmospheric variables can take on any of a continuum of values. Temperature, precipitation amount, geopotential height, windspeed, and other quantities are at least conceptually not restricted to integer values of the physical units in which they are measured. Even though the nature of measurement and reporting systems is such that atmospheric measurements are rounded to discrete values, the set of reportable values is usually large enough that most variables can still be treated as continuous quantities.

Many continuous theoretical distributions exist. Those used most frequently in the atmospheric sciences are discussed below. Encyclopedic information on these and many other continuous distributions can be found in Johnson and Kotz (1970a,b, 1972).

4.4.1 Distribution Functions and Expected Values

The mathematics of probability for continuous variables is somewhat different, although analogous, to that for discrete random variables. Whereas probability

calculations for discrete distributions involve summation over a discontinuous probability distribution function [e.g., Eqs. (4.1) and (4.6)], probability calculations for continuous random variables involve integration over continuous functions called *probability density functions*, or (PDFs).

Conventionally, the PDF for a random variable X is denoted $f(x)$. Just as a summation of a discrete probability distribution function over all possible values of the random quantity must equal 1, the integral of any PDF over all allowable values of x must equal 1:

$$\int_x f(x)\ dx = 1. \tag{4.11}$$

A function cannot be a PDF unless it satisfies this equation. Furthermore, $f(x)$ must be nonnegative for all values of x. No specific limits of integration have been included in Eq. (4.11), because different probability densities are defined over different ranges of the random variable.

Probability density functions are the continuous, theoretical, analogs of the familiar data histogram. However, the meaning of the PDF is often initially confusing precisely because of this analogy. In particular, the height of the density function $f(x)$, obtained when it is evaluated at a particular value of the random variable, is not in itself meaningful in a probability sense. The confusion arises because it is seldom realized that probability is proportional to area, and not to height, in both the PDF and the histogram.

Figure 4.2 shows a hypothetical PDF, defined on nonnegative values of a random variable X. While a probability density function can be evaluated for specific values of the random variable, say $X = 1$, by itself $f(1)$ is not meaningful in terms of probabilities for X. In fact, since X varies continuously over some segment of

Fig. 4.2 Hypothetical PDF $f(x)$ for a nonnegative random variable, X. Evaluation of $f(x)$ is not, by itself, meaningful in terms of probabilities for specific values of X. Probabilities are obtained by integrating portions of $f(x)$.

the real numbers, the probability of *exactly* $X = 1$ is infinitesimally small. It is meaningful, however, to think about and compute probabilities for values of a random variable in noninfinitesimal neighborhoods around $X = 1$. Figure 4.2 shows the probability of X between 0.5 and 1.5 as the integral of the PDF between these limits.

An idea related to the PDF is that of the *cumulative distribution function* (CDF). The CDF is a function of the random variable X, given by the integral of the PDF up to a particular value of x. Thus, the CDF specifies probabilities that the random quantity X will not exceed particular values. It is therefore the continuous, theoretical counterpart to the empirical CDF, Eq. (3.12). Conventionally, CDFs are denoted $F(x)$:

$$F(x) = \Pr\{X \leq x\} = \int_{X \leq x} f(x)\, dx. \tag{4.12}$$

Again, specific integration limits have been omitted from Eq. (4.12) to indicate that the integration is performed from the minimum allowable value of X to the particular value, x, that is the argument of the function. Since the values of $F(x)$ are probabilities, $0 \leq F(x) \leq 1$. Analogous functions, also called CDFs, can be obtained for discrete random variables by summation rather than integration.

Equation (4.12) transforms a particular value of the random variable to a cumulative probability. The value of the random variable corresponding to a particular cumulative probability is given by the inverse of the CDF

$$F^{-1}(p) = x(F), \tag{4.13}$$

where p is the cumulative probability. That is, Eq. (4.13) specifies the upper limit of the integration in Eq. (4.12) that will yield a particular cumulative probability $p = F(x)$. Since this inverse of the CDF specifies the data quantile corresponding to a particular probability, Eq. (4.13) is sometimes called the *quantile function.* Depending on the theoretical distribution being used, it may or may not be possible to write an explicit formula for the CDF or its inverse.

Statistical expectations are also defined for continuous random variables. As is the case for discrete variables, the expected value of a variable or a function is the probability-weighted average of that variable or function. Since probabilities for continuous random variables are computed by integrating their density functions, the expected value of a function of a random variable is given by the integral

$$E[g(x)] = \int_x g(x)\, f(x)\, dx. \tag{4.14}$$

Expectations of continuous random variables exhibit the properties listed in Eq. (4.9). For $g(x) = x$, $E[X] = \mu$ is the mean of the distribution whose density function is $f(x)$. Similarly, the variance of a continuous variable is given by the expectation of the function $g(x) = (x - E[X])^2$,

$$\begin{aligned}\mathrm{Var}[X] = E[(x - E[X])^2] &= \int_x (x - E[X])^2 f(x)\, dx \\ &= \int_x x^2 f(x)\, dx - (E[X])^2 \\ &= E[X^2] - \mu^2. \end{aligned} \tag{4.15}$$

Note that, depending on the particular functional form of $f(x)$, some or all of the integrals in Eqs. (4.12), (4.14), and (4.15) may not be analytically computable, and for some distributions the integrals may not even exist.

4.4.2 Gaussian Distribution

The Gaussian distribution plays a central role in classical statistics, and has many applications in the atmospheric sciences as well. It is sometimes also called the *normal distribution,* although this name carries the unwanted connotation that it is in some way universal, or that deviations from it are in some way unnatural. Its PDF is the familiar bell-shaped curve familiar even to people who have not studied statistics.

The breadth of applicability of the Gaussian distribution follows in large part from a very powerful theoretical result, known as the *central limit theorem.* Informally, the central limit theorem states that in the limit, as the sample size becomes large, the sum (or, equivalently, the arithmetic mean) of a set of independent observations will have a Gaussian distribution. This is true regardless of the distribution from which the observations have been drawn. The observations need not even be from the same distribution! Actually, the independence of the observations is not really necessary, either, which implies that the central limit theorem is directly applicable to atmospheric data.

What is not clear for particular data sets is just how large the sample size must be for the central limit theorem to apply. In practice this sample size depends on the distribution from which the summands are drawn. If the summed observations are themselves taken from a Gaussian distribution, the sum of any number of them (including, of course, $n = 1$) will also be Gaussian. For underlying distributions not too unlike the Gaussian (unimodal and not too asymmetrical), the sum of a modest number of observations will be nearly Gaussian. Summing daily temperatures to obtain a monthly-averaged temperature is a good example of this situation. Daily temperature values can exhibit noticeable asymmetry (e.g., Fig. 3.5), but are usually much more symmetrical than daily precipitation values. Conventionally, average daily temperature is approximated as the average of the maximum and minimum temperatures, so that the average monthly temperature is computed as

$$\bar{T} = \frac{1}{30} \sum_{i=1}^{30} \frac{T_{\max}(i) + T_{\min}(i)}{2}, \tag{4.16}$$

for a month with 30 days. Here the average monthly temperature derives from the sum of 60 numbers drawn from two more-or-less symmetrical distributions. It is

not surprising, in light of the central limit theorem, that monthly temperature values are often very successfully represented by Gaussian distributions.

A contrasting situation is that of the monthly total precipitation, constructed as the sum of, say, 30 daily precipitation values. There are fewer numbers going into this sum than is the case for the average monthly temperature in Eq. (4.16), but the more important difference has to do with the distribution of the underlying daily precipitation amounts. Typically most daily precipitation values are zero, and most of the nonzero amounts are small. That is, distributions of daily precipitation amounts are usually very strongly skewed to the right. Generally, the distribution of sums of 30 such values is also skewed to the right, although not so extremely. The schematic plot for $\lambda = 1$ in Fig. 3.9 illustrates this asymmetry for total January precipitation at Ithaca. In humid climates, the distributions of seasonal (i.e., 90-day) precipitation totals begin to approach the Gaussian, but even annual precipitation totals at arid locations can exhibit substantial positive skewness.

The probability density function for the Gaussian distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[- \frac{(x - \mu)^2}{2\sigma^2} \right]. \tag{4.17}$$

The two distribution parameters are the mean, μ, and the standard deviation, σ; while π is the mathematical constant $3.14159 \cdots$. Gaussian random variables are defined on the entire real line, so Eq. (4.17) is valid for $-\infty < x < \infty$. Graphing Eq. (4.17) results in the familiar bell-shaped curve shown in Fig. 4.3. This figure

Fig. 4.3 Probability density function for the Gaussian distribution. The mean, μ, locates the center of this symmetrical distribution, and the standard deviation, σ, controls the degree to which the distribution spreads out. Nearly all the probability is within $\pm 3\sigma$ of the mean.

shows that the mean locates the center of this symmetrical distribution, and the standard deviation controls the degree to which the distribution spreads out. Nearly all the probability is within $\pm 3\sigma$ of the mean.

In order to use the Gaussian distribution to represent a set of data, it is necessary to fit the two distribution parameters to the data. Good parameter estimates for this distribution are easily obtained using the method of moments. Again, the method of moments amounts to nothing more than equating the sample moments and the distribution, or "population," moments. The first moment is the mean, μ, and the second moment is the variance, σ^2. Therefore, one simply estimates μ as the sample mean [Eq. (1.1)], and σ as the sample standard deviation [Eq. (3.5)].

If a data sample follows approximately a Gaussian distribution, these parameter estimates will make Eq. (4.17) behave similarly to the data. Then, in principle, probabilities for events of interest are obtained by integrating Eq. (4.17). Practically, however, analytic integration of Eq. (4.17) is impossible, so that a formula for the CDF, $F(x)$, for the Gaussian distribution does not exist. Rather, Gaussian probabilities are obtained in one of two ways. If the probabilities are needed as part of a computer program, the integral of Eq. (4.17) can be economically approximated (e.g., Abramowitz and Stegun, 1984) to precision that is more than adequate. If only a few probabilities are needed, it is practical to compute them by hand using tabulated values such as those in Table B.1 in Appendix B.

In either of these two situations, a data transformation will nearly always be required. This is because Gaussian probability tables and algorithms pertain to the "standard" Gaussian distribution: the Gaussian distribution having $\mu = 0$ and $\sigma = 1$. Conventionally, the random variable described by the standard Gaussian distribution is denoted Z. Its probability density simplifies to

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2} \right]. \tag{4.18}$$

Any Gaussian random variable, X, can be transformed to standard form, Z, simply by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{X - \mu}{\sigma}. \tag{4.19}$$

In practical settings, the mean and the standard deviation usually need to be estimated using their sample estimates, so that one uses

$$z = \frac{x - \bar{x}}{s}. \tag{4.20}$$

Note that whatever physical units characterize x cancel in this transformation, so that the standardized variable, z, is always dimensionless.

Equation (4.20) is exactly the same as the "standardized anomaly" of Eq. (3.16). Any batch of data can be transformed by subtracting the mean and

dividing by the standard deviation, and this transformation will produce a transformed batch of data having a sample mean of zero and a sample standard deviation of one. However, the transformed data will not follow a Gaussian distribution unless the untransformed variable does. The use of the standardized variable in Eq. (4.20) to obtain probabilities is illustrated in the following example.

Example 4.5. Evaluating Gaussian Probabilities

Consider a Gaussian distribution characterized by $\mu = 22.2°\,\mathrm{F}$ and $\sigma = 4.4°\,\mathrm{F}$. These values were fit to a set of average January temperatures at Ithaca. Suppose that one is interested in evaluating the probability that an arbitrarily selected January will have average temperature as cold as or colder than 21.4°F, the value observed in 1987 (Table A.1). Transforming this temperature using the standardization in Eq. (4.20) yields $z = (21.4°\,\mathrm{F} - 22.2°\,\mathrm{F})/4.4°\,\mathrm{F} = -0.18$. Thus the probability of a temperature as cold as or colder than 21.4°F is the same as the probability of a value of Z as small as or smaller than -0.18: $\Pr\{X \leq 21.4°\,\mathrm{F}\} = \Pr\{Z \leq -0.18\}$.

Evaluating $\Pr\{Z \leq -0.18\}$ is easy, using Table B.1 in Appendix B, which contains cumulative probabilities for the standard Gaussian distribution. This cumulative distribution function is so commonly used that it is conventionally given its own symbol, $\Phi(z)$, rather than $F(z)$, as would be expected from Eq. (4.12). Looking across the row labeled -0.1 to the column labeled 0.08 yields the desired probability, 0.4286. Evidently, there is a rather substantial probability that an average temperature this cold or colder will occur in January at Ithaca.

Notice that Table B.1 contains no rows for positive values of z. These are not necessary because the Gaussian distribution is symmetrical. This means, for example, that $\Pr\{Z \geq +0.18\} = \Pr\{Z \leq -0.18\}$, since there will be equal areas under the curve in Fig. 4.3 to the left of $Z = -0.18$, and to the right of $Z = +0.18$. Therefore, Table B.1 can be used more generally to evaluate probabilities for $Z > 0$ by applying the relationship

$$\Pr\{Z \leq z\} = 1 - \Pr\{Z \leq -z\}, \tag{4.21}$$

which follows from the fact that the total area under the curve of any probability density is 1 [Eq. (4.11)].

Using Eq. (4.21) it is straightforward to evaluate $\Pr\{Z \leq +0.18\} = 1 - 0.4286 = 0.5714$. The average January temperature at Ithaca to which $Z = +0.18$ corresponds is obtained by inverting Eq. (4.20):

$$x = sz + \bar{x}. \tag{4.22}$$

The probability is 0.574 that an average January temperature at Ithaca will be no greater than $(4.4°\,\mathrm{F})(0.18) + 22.2°\,\mathrm{F} = 26.8°\,\mathrm{F}$.

It is only slightly more complicated to compute probabilities for outcomes between two specific values, say, Ithaca January temperatures between 20 and 25°F.

Since the event $\{T \leq 20°\text{F}\}$ is a subset of the event $\{T \leq 25°\text{F}\}$, the desired probability, $\Pr\{20°\text{F} < T \leq 25°\text{F}\}$, can be obtained by the subtraction $\Phi(z_{25}) - \Phi(z_{20})$. Here $z_{25} = (25.0°\text{F} - 22.2°\text{F})/4.4°\text{F} = 0.64$, and $z_{20} = (20.0°\text{F} - 22.2°\text{F})/4.4°\text{F} = -0.50$. Therefore (from Table B.1), $\Pr\{20°\text{F} < T \leq 25°\text{F}\} = \Phi(z_{25}) - \Phi(z_{20}) = 0.739 - 0.480 = 0.259$.

Also occasionally required is evaluation of the inverse of the standard Gaussian CDF, $\Phi^{-1}(p)$. This function specifies values of the standard Gaussian variate, z, corresponding to particular cumulative probabilities, p. Again, an explicit formula for this function cannot be written, so evaluating Φ^{-1} requires using Table B.1 in reverse. To find the average January Ithaca temperature defining the lowest decile (i.e., the coldest 10% of Januaries), the body of Table B.1 would be searched for $\Phi(z) = 0.10$. This cumulative probability corresponds almost exactly to $z = -1.28$. Using Eq. (4.22), $z = -1.28$ corresponds to a January temperature of $(4.4°\text{F})(-1.28) + 22.2°\text{F} = 16.6°\text{F}$. ¤

Example 4.6. Gaussian Distribution in Operational Climatology

The Gaussian distribution has an important application in the operational reporting of monthly and seasonal temperature outcomes by the Climate Analysis Center of the U.S. National Weather Service (Ropelewski *et al.*, 1985). Figure 4.4, taken from Arkin (1989), is an example of this application. This figure shows average temperatures for January 1989 over the conterminous United States, expressed as quantiles of the local Gaussian distributions. That is, different values

Fig. 4.4 Average temperatures for January 1989 over the conterminous United States, expressed as percentile values of local Gaussian distributions. Florida and most of the midwest were substantially warmer than usual, while only a small portion of the western United States was cooler than usual. [From Arkin (1989), reproduced with permission of the American Meteorological Society.]

of μ and σ have been estimated for each location whose data are summarized in the map.

Maps in this format show graphically where the average temperature for a given location and month falls with respect to its own climatological distribution. Five temperature categories are mapped: less than the 10th percentile, between the 10th and 30th percentiles, between the 30th and 70th percentiles (this category does not appear in the legend, but occurred in the unshaded portions of the map), between the 70th and 90th percentiles, and warmer than the 90th percentile. The boundaries between these percentile values correspond to standard Gaussian variates, z, of -1.28, -0.53, $+0.53$, and $+1.28$, respectively. It is readily apparent that Florida and much of the midwest were substantially warmer than usual, with temperatures above the (locally defined) 90th percentile. In these areas, January 1989 was warmer than approximately 9 out of 10 Januaries in the climatological base period. Only two small areas in the western United States show temperatures colder than the 10th percentile for this month. That is, these locations were colder than approximately 9 out of 10 of the years in the climatological reference period.

The mapped values in Fig. 4.4 were arrived at by comparing the observed temperatures for January 1989 at a large number of locations with Gaussian distributions representing average January temperatures at these same locations. Each of these local Gaussian distributions was fit using observed temperature data for the 30 years 1951–1980. The result of a similar analysis of climatological temperatures is shown in Fig. 4.5, which shows Gaussian distribution parameters for average January temperature, computed using data from the years 1931–1984 (Cayan *et al.*, 1986). The Climate Analysis Center has chosen to represent the climatological January temperatures with a parametric distribution in order to smooth the data, and to simplify the automation of producing the map product by condensing the data to two parameters (a Gaussian mean and a standard deviation) per location. The Gaussian distribution is a logical choice for this application because of the central limit theorem and the nature of temperature measurements comprising the monthly averages, as indicated previously.

It is worth thinking about why the format shown in Fig. 4.5 has been chosen to display the monthly temperature outcomes. Mapping the actual average temperatures would give no indication of how unusual the temperatures during the month in question may have been, or even whether the month was a relatively cool or warm one at a particular location. An alternative approach to displaying the temperature outcomes would be to map the monthly temperature anomalies, specifically, the actual average temperatures minus the corresponding means. This approach was taken by the Climate Analysis Center until 1985. These temperature anomalies indicated clearly by their sign whether the month was warmer or colder than the long-term mean at different locations, and the absolute value of the anomaly gave some indication of the degree of unusualness of the temperature for that month.

Fig. 4.5 Long-term average January temperatures (°C), (a) and corresponding standard deviations (°C × 10), (b). These parameters have been computed using temperature data for the years 1931–1984. Together the two figures define Gaussian distributions representing January temperatures throughout the conterminous United States. January is clearly milder in the south and extreme west coasts, whereas the year-to-year variability of average January temperature is greatest in the northern high plains and interior northwest. [From Cayan *et al.* (1986).]

However, as can be seen from Fig. 4.5b, temperatures at some locations are intrinsically more variable than at others. Average January temperatures 3°C warmer or colder than the long-term mean in northern Montana are not at all unusual, which is clear from the standard deviations near 5°C. (Using Table B.1, $\Pr\{Z \leq -3/5\} = 0.274$.) By contrast, the standard deviation of January temperature in central coastal California is less than 1.5°C, so an anomaly of −3°C would be quite unusual ($\Pr\{Z \leq -3/1.5\} = 0.023$). Thus, presenting (unstandardized) anomalies tends to exaggerate the importance of departures in high-variability regions, and mask the importance of anomalies in low-variability regions. Mapping the temperature outcomes as percentiles in Fig. 4.4 amounts to converting the observed temperatures to standardized anomaly form, and then expressing the standardized anomalies as probabilities of the Gaussian distribution. ¤

4.4.3 Gamma Distribution

Many atmospheric variables are distinctly asymmetrical, and skewed to the right. Often the skewness occurs when there is a physical limit on the left that is relatively near the range of the data. Common examples of this situation are precipitation amounts or windspeeds, which are physically constrained to be nonnegative. While it is mathematically possible to fit Gaussian distributions in such situations, the results are generally not useful. For example, the January 1933–1982 precipitation data in Table A.2 can be characterized by a sample mean of 1.96 in., and a sample standard deviation of 1.12 in. These two statistics are sufficient to fit a Gaussian distribution to these data, but applying this fitted distribution leads to nonsense. In particular, using Table B.1, one can compute the probability of negative precipitation as $\Pr\{Z \leq (0.00 - 1.96)/1.12\} = \Pr\{Z \leq -1.75\} = 0.040$. This computed probability is not especially large, but on the other hand it is not vanishingly small. The true probability is exactly zero: Observing negative precipitation is impossible.

There are a variety of continuous distributions that are bounded on the left by zero and positively skewed. One commonly used choice, used especially often for representing precipitation data, is the gamma distribution. The gamma distribution is defined by the PDF

$$f(x) = \frac{(x/\beta)^{\alpha-1} \exp(-x/\beta)}{\beta\Gamma(\alpha)}, \qquad x, \alpha, \beta > 0. \tag{4.23}$$

The two parameters of the distribution are α, the *shape parameter;* and β, the *scale parameter.* The quantity $\Gamma(\alpha)$ is a value of the standard mathematical function known as the *gamma function*, defined by the definite integral

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt. \tag{4.24}$$

Table 4.4

Values of the Gamma Function, $\Gamma(\alpha)$ [Eq. (4.24)], for $1.00 \leq \alpha \leq 1.99$

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	1.0000	0.9943	0.9888	0.9835	0.9784	0.9735	0.9687	0.9642	0.9597	0.9555
1.1	0.9514	0.9474	0.9436	0.9399	0.9364	0.9330	0.9298	0.9267	0.9237	0.9209
1.2	0.9182	0.9156	0.9131	0.9108	0.9085	0.9064	0.9044	0.9025	0.9007	0.8990
1.3	0.8975	0.8960	0.8946	0.8934	0.8922	0.8912	0.8902	0.8893	0.8885	0.8879
1.4	0.8873	0.8868	0.8864	0.8860	0.8858	0.8857	0.8856	0.8856	0.8857	0.8859
1.5	0.8862	0.8866	0.8870	0.8876	0.8882	0.8889	0.8896	0.8905	0.8914	0.8924
1.6	0.8935	0.8947	0.8959	0.8972	0.8986	0.9001	0.9017	0.9033	0.9050	0.9068
1.7	0.9086	0.9106	0.9126	0.9147	0.9168	0.9191	0.9214	0.9238	0.9262	0.9288
1.8	0.9314	0.9341	0.9368	0.9397	0.9426	0.9456	0.9487	0.9518	0.9551	0.9584
1.9	0.9618	0.9652	0.9688	0.9724	0.9761	0.9799	0.9837	0.9877	0.9917	0.9958

In general, the gamma function must be evaluated numerically (e.g., Abramowitz and Stegun, 1984; Press *et al.*, 1986) or approximated using tabulated values, such as those given in Table 4.4. The gamma function, Eq. (4.24), satisfies the factorial recurrence relationship

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \tag{4.25}$$

allowing Table 4.4 to be extended indefinitely. For example, $\Gamma(3.50) = (2.50)\Gamma(2.50) = (2.50)(1.50)\Gamma(1.50) = (2.50)(1.50)(0.8862) = 3.323$. Similarly, $\Gamma(4.50) = (3.50)\Gamma(3.50) = (3.50)(3.323) = 11.631$.

The PDF of the gamma distribution, Eq. (4.23), takes on a wide variety of shapes depending on the value of the shape parameter, α. As illustrated in Fig. 4.6,

Fig. 4.6 Gamma distribution density functions for four values of the shape parameter, α.

for $\alpha < 1$ the distribution is very strongly skewed to the right, with $f(x) \to \infty$ as $x \to 0$. For $\alpha = 1$ the function intersects the vertical axis at $1/\beta$ for $x = 0$, and this special case of the gamma distribution is called the *exponential distribution*. For $\alpha > 1$ the gamma distribution density function begins at the origin, $f(0) = 0$. Progressively larger values of α result in less skewness and a shifting of probability density to the right. For very large values of α the gamma distribution approaches the Gaussian distribution in form. The parameter α is always dimensionless.

The role of the scale parameter, β, effectively is to "stretch" or "squeeze" (i.e., to scale) the gamma density function to the right or left, depending on the overall magnitudes of the data values represented. Notice that the random quantity x in Eq. (4.23) is divided by β in both places where it appears. The scale parameter β has the same physical dimensions as x. As the distribution is stretched to the right by larger values of β, its height must drop in order to satisfy Eq. (4.11). Conversely, as the density is squeezed to the left, its height must rise. These adjustments in height are accomplished by the β in the denominator of Eq. (4.23).

The versatility in shape of the gamma distribution makes it an attractive candidate for representing precipitation data, and it is often used for this purpose. However, it is more difficult to work with than the Gaussian distribution, because obtaining good parameter estimates from particular batches of data is not as straightforward as for the Gaussian distribution. The simplest approach is to use the method of moments. Even here, however, there is a complication. This is that the two parameters for the gamma distribution do not correspond exactly to moments of the distribution, as was the case for the Gaussian. The mean of the gamma distribution is given by the product $\alpha\beta$, and the variance is $\alpha\beta^2$. Equating these expressions with the corresponding sample quantities yields a set of two equations in two unknowns, which can be solved to yield the moments estimators

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2} \tag{4.26a}$$

and

$$\hat{\beta} = \frac{s^2}{\bar{x}}. \tag{4.26b}$$

The moments estimators for the gamma distribution are not too bad for large values of the shape parameter, perhaps $\alpha > 10$, but can give very bad results for small values of α (Thom, 1958; Wilks, 1990). The moments estimators are said to be "inefficient," in a technical sense of not making maximum use of the information in a data set. The practical consequence of this inefficiency is that particular values of the parameters in Eq. (4.26) are erratic, or unnecessarily variable, from data sample to data sample.

A much better approach to parameter fitting for the gamma distribution is to

use the method of maximum likelihood. For many distributions, including the gamma distribution, maximum-likelihood fitting requires an iterative procedure that is really only practical using a computer. A later section in this chapter discusses the procedure with enough generality that readers having a moderate facility with computer programming will be able to fit this and other distributions using the method.

There are two approximations to the maximum likelihood estimators for the gamma distribution that are simple enough to compute by hand. Both employ the sample statistic

$$D = \ln(\bar{x}) - \frac{1}{n}\sum_{i=1}^{n} \ln(x_i), \tag{4.27}$$

which is the difference between the natural log of the sample mean, and the mean of the logs of the data. Equivalently, the sample statistic D is the difference between the logs of the arithmetic and geometric means. Notice that the sample mean and the standard deviation are not sufficient to compute the statistic D, since each datum must be used to compute the second term in Eq. (4.27).

The first of the two maximum-likelihood approximations for the gamma distribution is due to Thom (1958). The Thom estimator for the shape parameter is

$$\hat{\alpha} = \frac{1 + \sqrt{1 + 4D/3}}{4D}, \tag{4.28}$$

after which the scale parameter is obtained from

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}. \tag{4.29}$$

The second approach is a polynomial approximation to the shape parameter (Greenwood and Durand, 1960). One of two equations is used:

$$\hat{\alpha} = \frac{0.5000876 + 0.1648852D - 0.0544274D^2}{D}, \quad 0 \le D \le 0.5772, \tag{4.30a}$$

or

$$\hat{\alpha} = \frac{8.898919 + 9.059950D + 0.9775373D^2}{17.79728D + 11.968477D^2 + D^3}, \quad 0.5772 \le D \le 17.0, \tag{4.30b}$$

depending on the value of D. The scale parameter is again subsequently estimated using Eq. (4.29).

As was the case for the Gaussian distribution, the gamma density function is not analytically integrable. Gamma distribution probabilities must therefore be obtained either by computing approximations to the CDF [i.e., the integral of

Eq. (4.23)] or from tabulated probabilities. Formulas and computer routines for this purpose can be found in Abramowitz and Stegun (1984) and Press *et al.* (1986), respectively. A table of gamma distribution probabilities is included as Table B.2 in Appendix B.

In either case, gamma distribution probabilities will be available for the "standard" gamma distribution, with $\beta = 1$. Therefore, it is nearly always necessary to transform by rescaling the variable X of interest (characterized by a gamma distribution with arbitrary scale parameter β) to the variable

$$\xi = \frac{x}{\beta}, \tag{4.31}$$

that follows a gamma distribution with $\beta = 1$. The standard gamma variate ξ is dimensionless. The shape parameter, α, will be the same for both X and ξ. The procedure is analogous to the transformation to the standardized Gaussian variable, z, in Eq. (4.20).

Cumulative probabilities for the standard gamma distribution are given by a mathematical function known as the *incomplete gamma function*, $P(\alpha, \xi) = \Pr\{\Xi \leq \xi\} = F(\xi)$. It is this function that was used to compute the probabilities in Table B.2. The cumulative probabilities for the standard gamma distribution in Table B.2 are arranged in an inverse sense to the Gaussian probabilities in Table B.1. That is quantiles (transformed data values, ξ) of the distributions are presented in the body of the table, and cumulative probabilities are listed as the column headings. Different probabilities are obtained for different shape parameters, α, which appear in the first column.

Example 4.7. Evaluating Gamma Distribution Probabilities

Consider again the data for January precipitation at Ithaca during the 50 years 1933–1982. The average January precipitation for this period is 1.96 in., and the mean of the logarithms of the monthly precipitation totals is 0.5346, so Eq. (4.27) yields $D = 0.139$. Both Thom's method [Eq. (4.28)] and the Greenwood–Durand formula [Equation (4.30a)] yield $\alpha = 3.76$ in. and $\beta = 0.52$ in. By contrast, the moments estimators [Eq. (4.26)] yield $\alpha = 3.09$ in. and $\beta = 0.64$ in.

Adopting the approximate maximum-likelihood estimators, we can evaluate the unusualness of the January 1987 precipitation total at Ithaca with the aid of Table B.2. That is, if the climatological variations in Ithaca January precipitation are represented by the fitted gamma distribution with $\alpha = 3.76$ in. and $\beta = 0.52$ in., the cumulative probability corresponding to 3.15 in. (Table A.1) can be computed.

First, applying Eq. (4.31), we can obtain the standard gamma variate $\xi = 3.15$ in./0.52 in. $= 6.06$. Adopting $\alpha = 3.75$ as the closest tabulated value to the fitted $\alpha = 3.76$, we can see that $\xi = 6.06$ lies between the tabulated values for $F(5.214) = 0.80$ and $F(6.354) = 0.90$. Interpolation of yields $F(6.06) = 0.874$,

indicating that there is approximately one chance in eight for a January this wet or wetter to occur at Ithaca. The probability estimate could be refined slightly by interpolating between the rows for $\alpha = 3.75$ and $\alpha = 3.80$ to yield $F(6.06) = 0.873$, although this would probably not be worth the effort.

Table B.2 can also be used to invert the gamma CDF to find precipitation values corresponding to particular cumulative probabilities, $\xi = F^{-1}(p)$. Dimensional precipitation values are then recovered by reversing the transformation in Eq. (4.31). Consider estimation of the median January precipitation at Ithaca. This will correspond to the value of ξ satisfying $F(\xi) = 0.50$, which, in the row for $\alpha = 3.75$ in Table B.2, is 3.425. The corresponding dimensional precipitation amount is given by the product $\xi\beta = (3.425)(0.52 \text{ in.}) = 1.78$ in. By comparison, the sample median of the precipitation data in Table A.2 is 1.72 in. It is not surprising that the median is less than the mean of 1.96 in., since the distribution is positively skewed. The (perhaps surprising) implication of this comparison is that below-normal (i.e., below-average) precipitation is more likely than above-normal precipitation. ¤

Example 4.8. Gamma Distribution in Operational Climatology

The gamma distribution is used by the U.S. Climate Analysis Center to report monthly and seasonal precipitation amounts in the same way that the Gaussian distribution is used to report average temperatures. Figure 4.7, which is the com-

Fig. 4.7 Precipitation totals for January 1989 over the conterminous United States, expressed as percentile values of local gamma distributions. Portions of the east and west were drier than usual, while parts of the central portion of the country were wetter. [From Arkin (1989), reproduced with permission of the American Meteorological Society.]

panion map to Fig. 4.4, shows an example of this format for January 1989. As was the case for reporting temperatures, precipitation amounts for this month are not shown as accumulated depths, but rather as quantiles corresponding to local climatological gamma distributions. The same shading scheme is used in both Figs. 4.7 and 4.4, but the distributions from which the percentiles are computed are quite different.

As was the case for temperature, it is immediately clear which regions received substantially less, slightly less, about the same, slightly more, or substantially more precipitation in January 1989 as compared to the underlying climatological distribution. The shapes of these distributions vary widely, as can be seen in Fig. 4.8, from Wilks and Eggleston (1992). Comparison with Fig. 4.6 reveals that the distributions of January precipitation in much of the southwestern United States are very strongly skewed, whereas the corresponding distributions in much of the east, and in the Pacific northwest are much more symmetrical. (The corresponding scale parameters can be obtained from the mean monthly precipitation and the shape parameter using $\beta = \mu/\alpha$.) One advantage of expressing monthly precipitation amounts with respect to the climatological gamma distributions is that these very strong differences in the shapes of the precipitation climatologies

Fig. 4.8 Gamma distribution shape parameters for January precipitation over the conterminous United States. The distributions in the southwest are strongly skewed, while those for most locations in the east are much more symmetrical. The distributions were fit using data from the 30 years 1951–1980. [From Wilks and Eggleston (1992), reproduced with permission of the American Meteorological Society.]

Fig. 4.9 Illustration of the precipitation categories in Fig. 4.7 in terms of a gamma distribution density function with $\alpha = 2$. Outcomes drier than the 10th percentile lie to the left of $q_{0.1}$. Areas with precipitation between the 30th and 70th percentiles (between $q_{0.3}$ and $q_{0.7}$) would be unshaded on the map. Precipitation in the wettest 10% of the climatological distribution lie to the right of $q_{0.9}$.

do not confuse comparisons between locations. The advantages described above for reporting temperatures in terms of Gaussian distribution percentiles apply here as well.

Figure 4.9 illustrates the definition of the percentiles using a gamma PDF with $\alpha = 2$. The distribution is divided into five categories corresponding to the five shading levels in Fig. 4.7, with the precipitation amounts $q_{0.1}$, $q_{0.3}$, $q_{0.7}$, and $q_{0.9}$ separating regions of the distribution containing 10%, 20%, 40%, 20%, and 10% of the probability, respectively. As can be seen in Fig. 4.8, the shape of the distribution in Fig. 4.9 is characteristic of January precipitation for many locations in the midwestern United States and southern plains. For the stations in northeastern Oklahoma reporting January 1989 precipitation above the 90th percentile in Fig. 4.7, the corresponding precipitation amounts would have been larger than the locally defined $q_{0.9}$. ¤

4.4.4 Weibull Distribution

There are several distributions having in common with the gamma distribution the properties of nonnegativity and positive skewness. One of these that is commonly used in the atmospheric sciences is the Weibull distribution.

The Weibull distribution has been used most often in the atmospheric sciences to model variations in wind speeds. It has the PDF

$$f(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \qquad x, \alpha, \beta > 0. \tag{4.32}$$

Fig. 4.10 Weibull distribution probability density functions for four values of the shape parameter, α.

As is the case for the gamma distribution, the two parameters α and β are called the shape and scale parameters, respectively. The form of the Weibull distribution is also controlled similarly by the two parameters. The response of the shape of the distribution to different values of α is shown in Fig. 4.10. In common with the gamma distribution, $\alpha \leq 1$ produces reverse J shapes and strong positive skewness. Actually, the Weibull and gamma distributions are identical for $\alpha = 1$, for which both of these distributions reduce to the exponential distribution. For $\alpha = 3.6$ the Weibull is very similar to the Gaussian distribution. The scale parameter also acts similarly to either stretch or compress the basic shape along the x axis, for a given value of α.

One very convenient aspect of the Weibull distribution is that its density function is analytically integrable, so that a formula exists for the Weibull CDF:

$$F(x) = \Pr\{X \leq x\} = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]. \tag{4.33}$$

Tables of Weibull probabilities are therefore unnecessary.

Fitting parameters for the Weibull distribution requires iterative or other approximate methods. This is true even for the method of moments, since the distribution mean is given by $\mu = \beta\Gamma(1 + 1/\alpha)$ and its variance is $\sigma^2 = \beta^2[\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)]$. Since expressions for the moments estimators do not exist in closed form, using maximum likelihood to fit the parameters is the preferable approach. Both methods are discussed in Wong (1977). Some other estimation methods for Weibull parameters are given in Conradsen *et al.* (1984).

4.4.5 Lognormal Distribution

The lognormal distribution is another positively skewed, nonnegative distribution. Commonly it is used to represent variations in cloud properties, and it is also often used in hydrology. Fundamentally, applying the lognormal distribution amounts to nothing more than a logarithmic transformation of the data [i.e., Eq. (3.14a) with $\lambda = 0$], with the assumption that the resulting transformed data are described by a Gaussian distribution. That is, if the transformed variable $Y = \ln(X)$ follows the Gaussian distribution, Eq. (4.17), then the lognormal probability density for x is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[- \frac{(\ln x - \mu)^2}{2\sigma^2} \right] \tag{4.34}$$

where μ and σ are the mean and standard deviation, respectively, of the transformed variable, Y. Actually, the lognormal distribution is somewhat confusingly named, since it is the *anti*log of the random variable X that follows a Gaussian distribution.

Parameter fitting for the lognormal is simple and straightforward: The mean and standard deviation of the log-transformed data values y, i.e., μ_Y and σ_Y, respectively, are estimated by their sample counterparts. The relationships between these parameters in Eq. (4.34) and the mean and variance of the original variable X are

$$\mu_x = \exp\left[\mu_Y + \frac{\sigma_Y^2}{2} \right] \tag{4.35a}$$

and

$$\sigma_X^2 = (\exp[\sigma_Y^2] - 1) \exp[2\mu_Y + \sigma_Y^2]. \tag{4.35b}$$

Lognormal probabilities are evaluated simply by working with the transformed variable $y = \ln(x)$, and using computational routines or probability tables for the Gaussian distribution. In this case the Gaussian variable

$$Z = \frac{\ln(X) - \mu_Y}{\sigma_Y} \tag{4.36}$$

follows a Gaussian distribution with $\mu_z = 0$ and $\sigma_z = 1$.

4.4.6 Beta Distribution

Some variables of interest are restricted to segments of the real line that are bounded on two sides. Often such variables are restricted to the interval $0 \leq X \leq 1$. Examples of physically important variables subject to this restriction are cloud amount (observed as fraction of the sky) and relative humidity. An important, more abstract, variable of this type is probability, where a parametric distribution

can be useful in summarizing the frequency of use, say, of forecasts of daily rainfall probability. The usual choice to describe these types of variables is the beta distribution.

The PDF of the beta distribution is

$$f(x) = \left[\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\right] x^{p-1}(1-x)^{q-1},$$
$$0 \le x \le 1, \qquad p, q > 0. \tag{4.37}$$

This is a very flexible function, taking on many different shapes depending on the values of its two parameters, p and q. Figure 4.11 illustrates five of these. In general, for $p \le 1$ probability is concentrated near zero (e.g., $p = 0.25$ and $q = 2$, or $p = 1$ and $q = 2$, in Fig. 4.11), and for $q \le 1$ probability is concentrated near 1. If both parameters are less than one, the distribution is U-shaped. For $p > 1$ and $q > 1$ the distribution has a single mode ("hump") between 0 and 1 (e.g., $p = 2$ and $q = 4$, or $p = 10$ and $q = 2$, in Fig. 4.11), with more probability shifted to the right for $p > q$, and more probability shifted to the left for $q > p$. Beta distributions with $p = q$ are symmetrical. Reversing the values of p and q in Eq. (4.37) results in a density function that is the mirror image (horizontally flipped) of the original.

The beta distribution parameters are usually fit using the method of moments, although Johnson and Kotz (1970b) outline a trial-and-error procedure for maximum-likelihood estimation. Using the expressions for the first two moments of the distribution,

$$\mu = \frac{p}{p+q} \tag{4.38a}$$

Fig. 4.11 Five example PDFs for beta distributions. Mirror images of these distributions are obtained by reversing the parameters p and q.

and

$$\sigma^2 = \frac{pq}{(p + q)^2(p + q + 1)}, \tag{4.38b}$$

the moments estimators

$$\hat{p} = \frac{\bar{x}^2(1 - \bar{x})}{s^2} - \bar{x} \tag{4.39a}$$

and

$$\hat{q} = \frac{\hat{p}(1 - \bar{x})}{\bar{x}} \tag{4.39b}$$

are easily obtained.

Use of the beta distribution is not limited only to variables defined on the interval [0, 1]. A variable—say, Y—constrained to any interval $[a, b]$ can be represented by a beta distribution after subjecting it to the transformation

$$X = \frac{Y - a}{b - a}. \tag{4.40}$$

In this case parameter fitting is accomplished using

$$\bar{x} = \frac{\bar{y} - a}{b - a} \tag{4.41a}$$

and

$$s_X^2 = \frac{s_Y^2}{(b - a)^2}, \tag{4.41b}$$

which are then substituted into Eq. (4.39).

The integral of the beta probability density does not exist in closed form except for a few special cases. Probabilities can be obtained through numerical methods (Abramowitz and Stegun, 1984; Press *et al.*, 1986), where the CDF for the beta distribution is known as the *incomplete beta function*, $I_x(p, q) = \Pr\{0 \leq X \leq x\} = F(x)$. Tables of beta distribution probabilities are given in Epstein (1985) and Winkler (1972).

4.4.7 Gumbel Distribution

The statistical distributions of extreme atmospheric events can be of substantial interest for problems of engineering design, and for other risk assessments. The term *extreme event* indicates the largest (or smallest) value of an atmospheric variable among a given number of observations. Examples of extreme events could include the warmest maximum temperature observed at a location during a particular July, or the largest daily rainfall amount recorded at a location during a particular year.

Extreme-value data can be characterized using theoretical probability distributions in the same manner as other forms of data. Imagine, for example, that a data set consisting of the single warmest daily maximum temperatures from each of the last 30 Julys is assembled. These data would constitute a sample approximating the climatological distribution of the temperature of the hottest July day in a year. A theoretical distribution could be fit to these data, and then used to represent the climatological distribution of extreme July maximum temperatures. This theoretical distribution could be used, for example, to estimate probabilities that the hottest day in July of next (or any other) summer will be warmer than specified temperatures, or to estimate hottest-July-day temperatures corresponding to cumulative probabilities of, say, 0.99 or larger. These latter probabilities could not be obtained directly from the empirical distribution function of a set of data having $n = 30$.

Most of the distributions described previously in this chapter tend not to represent extreme-value data very well. A number of theoretical extreme value distributions exist (e.g., Buishand, 1989), but the most commonly used of these is the Gumbel distribution. The Gumbel distribution is so frequently used to represent the statistics of extremes that it is sometimes called *the* extreme-value distribution. It is also known as the *Fisher–Tippett type I* distribution. This distribution arises naturally as one of three possible distributions of the largest of a sequence of independent, identically distributed random variables (Gumbel, 1958).

The PDF for the Gumbel distribution is

$$f(x) = \frac{1}{\beta} \exp\left\{ -\exp\left[\frac{(x - \xi)}{\beta}\right] - \frac{(x - \xi)}{\beta}\right\}, \tag{4.42}$$

where ξ and β are location and scale parameters, respectively. This density function is skewed to the right, and exhibits its maximum at $x = \xi$. The PDF for the Gumbel distribution is analytically integrable, so that cumulative Gumbel distribution probabilities can be obtained from the CDF

$$F(x) = \exp\left\{ -\exp\left[- \frac{(x - \xi)}{\beta}\right]\right\}. \tag{4.43}$$

The simplest way to estimate the parameters of the Gumbel distribution is to use the method of moments. The moments estimators for the Gumbel distribution parameters are computed using the sample mean and standard deviation. The estimation equations are

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi} \tag{4.44a}$$

and

$$\hat{\xi} = \bar{x} - \gamma\hat{\beta}, \tag{4.44b}$$

where $\gamma = 0.57721 \cdots$ is Euler's constant.

4.5 Multivariate Probability Distributions

So far, this chapter has considered only univariate probability distributions—functions describing the variation of a single quantity. It is also possible to represent the joint distribution of two or more variables using parametric probability distributions. While some discrete theoretical multivariate distributions exist (Johnson and Kotz, 1969), usually continuous theoretical distributions are used in multivariate settings.

4.5.1 Bivariate Normal Distribution

By far the most extensively used continuous multivariate distribution is the generalization of the Gaussian to the joint distribution of two variables, known as the *bivariate normal* distribution. This distribution describes the joint behavior of two Gaussian variables, X and Y. It is sometimes possible to use this distribution to describe the behavior of two non-Gaussian variables if they are first subjected to the symmetry-producing transformations of Eqs. (3.14). In fact, the opportunity to use the bivariate normal can be a major motivation for using the power transformations.

The bivariate normal distribution is defined by the PDF

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \cdot \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]\right\}. \quad (4.45)$$

As a generalization of Eq. (4.17) from one to two dimensions, this function defines a surface above the X–Y plane rather than a curve above the X axis. For bivariate distributions, probability corresponds geometrically to the volume under this surface, so that analogously to Eq. (4.11), a necessary condition to be fulfilled by any bivariate PDF is

$$\int_x \int_y f(x, y)\, dy\, dx = 1, \qquad f(x, y) \geq 0. \quad (4.46)$$

The bivariate normal distribution has five parameters: the two means and standard deviations for the variables X and Y, and the correlation between them, ρ. The two marginal distributions for the variables X and Y [i.e., the univariate PDFs $f(x)$ and $f(y)$] must both be Gaussian distributions. These two marginal distributions have parameters μ_X, σ_X, and μ_Y, σ_Y, respectively. Fitting the bivariate normal distribution is very easy. These four parameters are estimated using their sample counterparts for the X and Y variables separately, and the parameter ρ is estimated as the Pearson product-moment correlation between X and Y, Eq. (3.17).

Fig. 4.12 Perspective view of a bivariate normal distribution with $\sigma_X = \sigma_Y$, and $\rho = -0.75$. The individual lines depicting the hump of the bivariate distribution have the shape of the (univariate) Gaussian distribution, illustrating that conditional distributions of X given a particular value of Y are themselves Gaussian.

Figure 4.12 illustrates the general shape of the bivariate normal distribution. It is mound-shaped in three dimensions, with properties that depend on the five parameters. The function achieves its maximum height above the point (μ_X, μ_Y). Increasing σ_X stretches the density in the X direction and increasing σ_Y stretches it in the Y direction. For equal standard deviations and $\rho = 0$ the density is axially symmetrical around the point (μ_X, μ_Y), and curves of constant height are concentric circles. As ρ increases in absolute value the density function is stretched diagonally, with the lines of constant height becoming increasingly elongated ellipses. For negative ρ the orientation of these ellipses is as depicted in Fig. 4.12: Larger values of X are more likely with smaller values of Y, and smaller values of X are more likely with larger values of Y. The ellipses have the opposite orientation (positive slope) for positive values of ρ.

Probabilities for joint outcomes of X and Y are given by the double integral of Eq. (4.45) over the relevant region in the plane, for example

$$\Pr\{(y_1 < Y \leq y_2) \cap (x_1 < X \leq x_2)\} = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)\, dy\, dx. \quad (4.47)$$

This integration cannot be done analytically, and in practice numerical methods are usually used. Probability tables for the bivariate normal do exist (National Bureau of Standards, 1959), but they are lengthy and cumbersome. It is possible to compute probabilities for elliptically shaped regions centered on (μ_X, μ_Y) using the method illustrated in Example 9.4. When computing probabilities for other regions, it can be more convenient to work with the bivariate normal distribution in standardized form. This is the extension of the standardized univariate Gaussian distribution [Eq. (4.18)], and is achieved by subjecting both the X and Y variables

to the normalizing transformation, Eq. (4.19). Thus $\mu_{Z_X} = \mu_{Z_Y} = 0$ and $\sigma_{Z_X} = \sigma_{Z_Y} = 1$, leading to the bivariate density

$$f(z_X, z_Y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left[- \frac{z_X^2 - 2\rho z_X z_Y + z_Y^2}{2(1 - \rho^2)} \right]. \tag{4.48}$$

A very useful property of the bivariate normal distribution is that the conditional distribution of one of the variables, given any particular value of the other, is Gaussian. This is illustrated graphically in Fig. 4.12, where the shapes of the individual lines defining the shape of the distribution in three dimensions have Gaussian shapes. Each indicates a function proportional to a conditional distribution of X given a particular value of Y. The parameters for these conditional Gaussian distributions can be calculated from the five parameters of the bivariate normal. For the distribution of X given a particular value of Y, the conditional Gaussian density function $f(x|Y = y)$ has parameters

$$\mu_{X|Y} = \mu_X + \rho\sigma_X \frac{(y - \mu_Y)}{\sigma_Y} \tag{4.49a}$$

and

$$\sigma_{X|Y} = \sigma_X\sqrt{1 - \rho^2}. \tag{4.49b}$$

Equation (4.49a) relates the mean of X to the standardized anomaly of Y. It indicates that the conditional mean $\mu_{X|Y}$ is larger than the unconditional mean μ_X if Y is greater than its mean and ρ is positive, or if Y is less than its mean and ρ is negative. If X and Y are uncorrelated, knowledge of a value of Y gives no additional information about X, and $\mu_{X|Y} = \mu_X$ since $\rho = 0$. Equation (4.49b) indicates that, unless the two variables are uncorrelated, $\sigma_{X|Y} < \sigma_X$, regardless of the sign of ρ. Here, knowledge of Y provides some information about X, and the diminished uncertainty about X is reflected in a smaller standard deviation.

Example 4.9. Bivariate Normal Distribution and Conditional Probability

Consider the maximum temperature data for January 1987 at Ithaca and Canandaigua, in Table A.1. Figure 3.5 indicates that these data are fairly symmetrical, so that it may be reasonable to model their joint behavior as bivariate normal. A scatterplot of these two variables is shown in one of the panels of Fig. 3.18. The average maximum temperatures are 29.87 and 31.77°F at Ithaca and Canandaigua, respectively. The corresponding sample standard deviations are 7.71 and 7.86°F. Table 3.3 shows their Pearson correlation to be 0.957.

With such a high correlation, knowledge of the temperature at one location should give very strong information about the temperature at the other. Suppose it is known that the Ithaca maximum temperature is 25°F and probability information about the Canandaigua maximum temperature is desired. Using Eq. (4.49a), the conditional mean for the distribution of maximum temperature at

Fig. 4.13 Gaussian distributions, representing the unconditional distribution for daily January maximum temperature at Canandaigua, and the conditional distribution given that the Ithaca maximum temperature was 25°F. The high correlation between maximum temperature at the two locations results in the conditional distribution being much sharper, reflecting substantially diminished uncertainty.

Canandaigua, given that the Ithaca maximum temperature is 25°F, is 27.1°F—substantially lower than the unconditional mean of 31.77°F. Using Eq. (4.49b), the conditional standard deviation is 2.28°F. [This would be the conditional standard deviation regardless of the particular value of the Ithaca temperature chosen, since Eq. (4.49b) does not depend on the value of the conditioning variable.] The conditional standard deviation is so much lower than the unconditional standard deviation because of the high correlation of maximum temperature between the two locations. As illustrated in Fig. 4.13, this reduced uncertainty means that any conditional distribution for Canandaigua temperature given the Ithaca temperature will be much sharper than the unmodified, unconditional distribution for Canandaigua maximum temperature.

Using these parameters for the conditional distribution of maximum temperature at Canandaigua, one can compute such quantities as the probability that the Canandaigua maximum temperature is at or below freezing, given that the Ithaca maximum is 25°F. The required standardized variable is $z = (32 - 27.1)/2.28 = 2.14$, which corresponds to a probability of 0.984. By contrast, the corresponding climatological probability (without benefit of knowing the Ithaca maximum temperature) would be computed from $z = (32 - 31.8)/7.86 = 0.025$, corresponding to the much lower probability 0.510. ¤

4.5.2 Multivariate Normal Distribution

The Gaussian distribution generalizes easily to more than two dimensions as well. In the general case, the joint distribution of $K > 2$ Gaussian variables is described

by the multivariate normal distribution. It is impractical to work with the multivariate normal distribution without using matrix notation (Section 9.2). With this notation, the PDF for the multivariate normal is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det[\boldsymbol{\Sigma}]}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T[\boldsymbol{\Sigma}]^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right], \quad (4.50)$$

where $\mathbf{x}$ and $\boldsymbol{\mu}$ are k-dimensional vectors, $[\boldsymbol{\Sigma}]$ is the $(K \times K)$ variance–covariance matrix (Chapter 9) of the K variables in the vector $\mathbf{x}$ and det $[\boldsymbol{\Sigma}]$ is its determinant. The determinant is a scalar function of a matrix, that loosely corresponds to the idea of a "magnitude" for that matrix. If each of the K variables in $\mathbf{x}$ are separately standardized according to (4.19), the standardized multivariate normal density,

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det[\mathbf{R}]}} \exp\left[\frac{-\mathbf{z}^T[\mathbf{R}]^{-1}\mathbf{z}}{2} \right] \quad (4.51)$$

results, where $[\mathbf{R}]$ is the (Pearson) correlation matrix (e.g., Fig. 3.17) for the K variables.

A number of useful properties of the relationships among a suite of variables can be derived if multivariate normality can be justified or assumed. These properties are described in standard textbooks on multivariate statistics (e.g., Johnson and Wichern, 1982; Morrison, 1967), which are essentially entirely devoted to the multivariate normal distribution. One of these properties is that the conditional distributions of subsets of the vector $\mathbf{X}$, given particular values of the other elements, are also multivariate normal. This is the multivariate generalization of Eq. (4.49). The conditional mean of the subset of the variables $\mathbf{X}_1$ given particular values for the remaining variables $\mathbf{X}_2 = \mathbf{x}_2$ is

$$\boldsymbol{\mu}_1 \mid \mathbf{x}_2 = \boldsymbol{\mu}_1 + [\boldsymbol{\Sigma}_{12}][\boldsymbol{\Sigma}_{22}]^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \quad (4.52a)$$

and the conditional covariance matrix is

$$[\boldsymbol{\Sigma}_1 \mid \mathbf{x}_2] = [\boldsymbol{\Sigma}_{11}] - [\boldsymbol{\Sigma}_{12}]\,[\boldsymbol{\Sigma}_{22}]^{-1}[\boldsymbol{\Sigma}_{21}], \quad (4.52b)$$

where $[\boldsymbol{\Sigma}_{11}]$ is the covariance matrix for the variables in $\mathbf{X}_1$, $[\boldsymbol{\Sigma}_{22}]$ is the covariance matrix for the variables in $\mathbf{X}_2$, and $[\boldsymbol{\Sigma}_{12}] = [\boldsymbol{\Sigma}_{21}]^T$ is the matrix of covariances between the variables in $\mathbf{X}_1$ and $\mathbf{X}_2$. These four matrices comprise the covariance matrix for both sets of variables jointly:

$$[\boldsymbol{\Sigma}] = \mathrm{Var}[\mathbf{X}] = \mathrm{Var}\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} [\boldsymbol{\Sigma}_{11}] & [\boldsymbol{\Sigma}_{12}] \\ [\boldsymbol{\Sigma}_{21}] & [\boldsymbol{\Sigma}_{22}] \end{bmatrix}. \quad (4.53)$$

An implication of Eq. (4.52) is that the univariate, or marginal, distributions of each of the K variables will be Gaussian. However, it is not necessarily the case that the joint distribution of an arbitrarily selected set of K Gaussian variables will follow the multivariate normal distribution.

4.6 Qualitative Assessments of the Goodness of Fit

Having fit a theoretical distribution to a batch of data, it is of more than passing interest to verify that the theoretical probability model provides an adequate description. Fitting an inappropriate distribution can lead to erroneous conclusions being drawn from the theoretical model. Quantitative methods for evaluating the closeness of fitted distributions to underlying data rely on ideas from formal hypothesis testing. A few such methods will be presented in the next chapter, which pertains to statistical tests. Given below are some qualitative, graphical methods useful for subjectively discerning the goodness of fit. These methods are useful and instructive even if a formal goodness-of-fit test of fit is to be computed. While a formal test may indicate an inadequate fit, it will not inform the data analyst as to the nature of the problem. Graphical comparisons of the data and the fitted distribution allow diagnosis of where and how the theoretical representation is inadequate.

4.6.1 Superposition of a Fitted Theoretical Distribution and Data Histogram

Probably the simplest and most intuitive means of comparing a fitted theoretical distribution to the underlying data is superposition of the fitted distribution and a histogram of the data. Gross departures from the data can be readily seen in this way and, if the data are sufficiently numerous, irregularities in the histogram due to sampling variations will not be too distracting.

For discrete distributions, the probability distribution function is already very much like the histogram. Both the histogram and the probability distribution function assign probability to a discrete set of outcomes. Comparing the two requires only that the same discrete data values, or ranges of the data values, be plotted, and that the histogram and distribution function be scaled comparably. This second condition is met by plotting the histogram in terms of relative, rather than absolute, frequency on the vertical axis. Figure 4.1 is an example of the superposition of a Poisson probability distribution function on the histogram of observed annual numbers of tornados in New York State.

The procedure for superimposing a continuous PDF on a histogram is entirely analogous. The fundamental constraint is that the integral of any PDF, over the full range of the random variable, must be one. That is, Eq. (4.11) is satisfied by all PDFs. One approach to matching the histogram and the density function is to rescale the density function. The proper scaling factor is obtained by computing the area occupied collectively by all the bars in the histogram plot. Denoting this area as A, it is easy to see that multiplying the fitted density function $f(x)$ by A produces a curve whose area is also A. This is because, as a constant, A can be taken out of the integral: $\int A \cdot f(x)\, dx = A \cdot \int f(x)\, dx = A \cdot 1 = A$. Note that it is also possible to rescale the histogram heights so that the total area contained in

Fig. 4.14 Histogram of the 1933–1982 Ithaca January precipitation data from Table A.2, superimposed with the fitted gamma distribution (solid line) and the fitted Gaussian distribution (broken line). Each of the two density functions has been multiplied by $A = 25$, since the bin width is 0.5 in. and there are 50 observations. Evidently the gamma distribution provides a reasonable representation of the data, whereas the Gaussian distribution underrepresents the right tail and implies nonzero probability for negative precipitation.

the bars is 1. This latter approach is more traditional in statistics, where the histogram is regarded as an estimate of the density function.

Example 4.10. Superposition of Density Functions onto a Histogram

Figure 4.14 illustrates the procedure of superimposing fitted distributions and a histogram for the 1933–1982 January precipitation totals at Ithaca from Table A.2. Here $n = 50$ years of data, and the bin width for the histogram [consistent with Eq. (3.11a)] is 0.5 in., so the area occupied by the histogram rectangles is $A = (50)(0.5) = 25$. Superimposed on this histogram are density functions of the gamma distribution fit using Eq. (4.28) or (4.30a) (solid curve) and the Gaussian distribution fit by matching the sample and distribution moments (dashed curve). In both cases the density functions [Eqs. (4.23) and (4.17), respectively] have been multiplied by 25 so that their areas are equal to that of the histogram. It is clear that the symmetrical Gaussian distribution is a poor choice for representing these positively skewed precipitation data, since too little probability is assigned to the largest precipitation amounts and nonnegligible probability is assigned to impossible negative precipitation amounts. The gamma distribution represents these data much more closely, and provides a quite plausible summary

of the year-to-year variations in the data. The fit appears to be worst for the 0.75–1.25-in. and 1.25–1.75-in. bins, although this could easily have resulted from sampling variations. If two or three data values had been shifted from the wetter to the drier category, the fit would look quite good, and there are seven observations between 1.30 and 1.36 in. This same data set will also be used in Chapter 5 to test formally the fit of these two distributions. ¤

4.6.2 Probability Plots

Probability plots compare empirical (data) and fitted CDFs. For the fitted theoretical distribution the link between values of the random variable and cumulative probability is made through the CDF. Cumulative probability corresponding to a particular value x is simply $F(x)$, and a value of x corresponding to a particular cumulative probability is $F^{-1}(p)$ [Eq. (4.13)]. The empirical CDF, estimated using expressions such as Eq. (3.13a) or (3.13b), that depend on the ranked data, plays the corresponding role for the observed data values.

The quantile–quantile (Q–Q) plot is a probability plot comparing the fitted and empirical cumulative distribution functions in terms of the dimensional values of the variable (the empirical quantiles). The Q–Q plot is a scatterplot. Each coordinate pair defining the location of a point consists of a data value and the corresponding theoretical estimate for that data value derived from the empirical cumulative probability estimate. Adopting Eq. (3.13a) as that estimator for cumulative probability, each plotted point in a Q–Q plot would have the Cartesian coordinates $(F^{-1}[i/(n+1)], x_{(i)})$. Thus the ith point on the Q–Q plot is defined by the ith smallest data value, $x_{(i)}$, and the value of the theoretical random variable corresponding to the sample cumulative probability $p = i/(n+1)$ in the fitted distribution.

Figure 4.15 shows Q–Q plots comparing the fits of four theoretical distributions to extreme rainfall data for Baltimore, Maryland. The Q–Q plots allow easy examination of the four fits, especially when the 1 : 1 lines have been included. A Q–Q plot for a distribution that represented the data perfectly would have all points falling on this line. All four distributions in Fig. 4.15 appear to do a good job in the central part of the data, as the points fall close to the 1 : 1 line here. Precipitation extreme data are particularly difficult to fit on the right tail. Once the diagrams have been constructed it is easy to see that the beta-P distribution represents the data quite well through its entire range. The generalized gamma (Gen. Γ) distribution allocates insufficient probability on the right tail, but too much probability on the left: Observed precipitation is substantially above the fitted precipitation estimates $F^{-1}[i/(n+1)]$ on both tails. The beta-κ also allocates insufficient probability to the extreme right tail, although in the lower part of the distribution the fit is quite close. Conversely, the Generalized Extreme Value (GEV) distribution allocates too much probability to the right tail, which can be seen from the fact that the points for the largest precipitation amounts lie below the 1 : 1 line.

Fig. 4.15 Quantile–quantile plots showing the fit of four theoretical distributions to the largest 91 daily precipitation observations at Baltimore, MD, for the years 1900–1990. Observed precipitation amounts are on the vertical, and amounts inferred from the fitted distributions are on the horizontal. Diagonal lines indicate 1 : 1 correspondence. [From Wilks (1993).]

Another type of probability plot is the P–P plot, which compares empirical and fitted distributions according to cumulative probabilities rather than dimensional data values. Again adopting Eq. (3.13a), each point plotted in a probability plot would have the Cartesian coordinates $(F[x_{(i)}], i/(n + 1))$. Conceptually, the Q–Q plot and the P–P plot bear an inverse relationship to each other. Again, a perfect representation of the data by the theoretical distribution results in the points of the probability plot falling on the 1 : 1 line.

Figure 4.16 shows two P–P plots, corresponding to the January 1933–1982 precipitation data and the two fitted PDFs shown in Fig. 4.14. It is apparent that the cumulative probabilities from the gamma distribution lie closer to the 1 : 1 line. That the points for the Gaussian distribution fall below the 1 : 1 line for smaller probabilities is a reflection of its assignment of positive probability to negative precipitation amounts. The jumps in these probability plots between about 0.25 and 0.35 in empirical probability correspond to the large histogram bar between 1.25 and 1.75-in. in Fig. 4.14. Points for the Gaussian distribution far above the line for most of the higher probabilities indicate insufficient probability allocation in this region. The apparently closer agreement between the Gaussian and empirical probabilities for the very highest data values points out that probability

Fig. 4.16 P–P plots for gamma (circles) and Gaussian (x's) distributions fit to the January 1933–1982 Ithaca precipitation data in Table A.2. Points for the gamma distribution are much closer to the 1 : 1 line, indicating a better fit to the data.

plots are relatively insensitive to differences in the extreme tails (again, compare Fig. 4.14). A Q–Q plot for these data using the Gaussian distribution would broadly resemble that for the generalized gamma (Gen. Γ) distribution in Fig. 4.15.

4.7 Parameter Fitting Using Maximum Likelihood

For many distributions, fitting parameters using the simple method of moments produces inferior results that can lead to misleading extrapolations and inferences. The method of maximum likelihood is a versatile and important alternative. As the name suggests, the method seeks to find values of the distribution parameters that maximize the *likelihood function.* The procedure follows from the notion that the likelihood is a measure of the degree to which the data support particular values of the parameter(s) (e.g., Lindgren, 1976). A Bayesian interpretation of the procedure (except for small sample sizes) would be that the maximum-likelihood

estimators are the most probable values for the parameters, given the observed data.

Notationally, the likelihood function for a single observation, x, looks identical to the probability density (or, for discrete variables, the probability distribution) function. The distinction is that the probability density is a function of the data for fixed values of the parameters, whereas the likelihood function is a function of the unknown parameters for fixed values of the (already observed) data. The likelihood function for the parameters of a distribution given a sample of n independent data values is just the product of the n individual likelihood functions. For example, the likelihood function for the Gaussian parameters μ and σ, given a sample of n observations, x_i, $i = 1, \ldots, n$, is

$$\mathfrak{l}(\mu, \sigma) = \sigma^{-n}(\sqrt{2\pi})^{-n} \prod_{i=1}^{n} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]. \tag{4.54}$$

Actually, the likelihood can be any function proportional to Eq. (4.54), so the constant factor involving the square root of 2π could have been omitted. It has been included to emphasize the relationship between Eqs. (4.54) and (4.17). Geometrically, Eq. (4.54) describes a surface above the μ–σ plane that takes on a maximum value above a specific pair of parameter values, depending on the particular data set given by the x_i values.

Usually it is more convenient to work with the logarithm of the likelihood function, known as the *log-likelihood.* Since the logarithm is a strictly increasing function, the same parameter values will maximize both the likelihood and log-likelihood functions. The log-likelihood function for the Gaussian parameters, corresponding to Eq. (4.54), is

$$\begin{aligned} L(\mu, \sigma) &= \ln[\mathfrak{l}(\mu, \sigma)] \\ &= -n \ln(\sigma) - n \ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2, \end{aligned} \tag{4.55}$$

where, again, the term involving 2π is not strictly necessary because it does not depend on the parameters μ and σ.

Conceptually, at least, maximizing the log-likelihood is a straightforward exercise in calculus. For the Gaussian distribution the exercise really is simple, since the maximization can be done analytically. Taking derivatives of Eq. (4.55) with respect to the parameters yields

$$\frac{\partial L(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} x_i - n\mu\right] \tag{4.56a}$$

and

$$\frac{\partial L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{4.56b}$$

Setting each of these derivatives equal to zero and solving yields, respectively,

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.57a}$$

and

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}. \tag{4.57b}$$

These are the maximum-likelihood estimators (MLEs) for the Gaussian distribution, which are readily recognized as being very similar to the moments estimators. The only difference is the divisor in Eq. (4.57b), which is n rather than $n - 1$. The divisor $n - 1$ is often adopted when computing the sample standard deviation, because that choice yields an unbiased estimate of the population value. This difference points out the fact that the MLEs for a particular distribution may not be unbiased.

The MLEs for the Gaussian distribution are somewhat unusual, in that they can be computed analytically. It is more typical for approximations to the MLEs to be computed iteratively. The most common approach is to think of the maximization of the log-likelihood as a nonlinear rootfinding problem to be solved using the multidimensional generalization of the Newton–Raphson method (e.g., Press *et al.*, 1986). This approach follows from the truncated Taylor expansion of the derivative of the log-likelihood function

$$L'(\theta^*) \approx L'(\theta) + (\theta^* - \theta)L''(\theta), \tag{4.58}$$

where θ denotes a generic vector of distribution parameters and θ^* are the true values to be approximated. Since it is the *derivative* of the log-likelihood function, $L'(\theta^*)$, whose roots are to be found, Eq. (4.58) requires computation of the second derivatives of the log-likelihood, $L''(\theta)$. Setting Eq. (4.58) equal to zero (to produce a maximum in the log-likelihood, L) and rearranging yields the expression describing the algorithm for the iterative procedure:

$$\theta^* = \theta - \frac{L'(\theta)}{L''(\theta)}. \tag{4.59}$$

Beginning with an initial guess, θ, one computes an updated set of estimates, θ^*, which are in turn used as the guesses for the next iteration.

Example 4.11. Algorithm for MLEs of Gamma Distribution Parameters

In practice, Eq. (4.59) is somewhat complicated by the fact that usually more than one parameter must be estimated, so that $L'(\theta)$ is a vector of first derivatives, and $L''(\theta)$ is a matrix of second derivatives. To illustrate, consider the gamma distribution [Eq. (4.23)]. For this distribution, Eq. (4.59) becomes

$$\begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \begin{bmatrix} \dfrac{\partial^2 L}{\partial \alpha^2} & \dfrac{\partial^2 L}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2 L}{\partial \beta \partial \alpha} & \dfrac{\partial^2 L}{\partial \beta^2} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\partial L}{\partial \alpha} \\ \dfrac{\partial L}{\partial \beta} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \begin{bmatrix} -n\Gamma''(\alpha) & \dfrac{-n}{\beta} \\ \dfrac{-n}{\beta} & \dfrac{n\alpha}{\beta^2} - \dfrac{2\Sigma x}{\beta^3} \end{bmatrix}^{-1}$$

$$\begin{bmatrix} \Sigma \ln(x) - n \ln(\beta) - n\Gamma'(\alpha) \\ \dfrac{\Sigma x}{\beta^2} - \dfrac{n\alpha}{\beta} \end{bmatrix}, \tag{4.60}$$

where $\Gamma'(\alpha)$ and $\Gamma''(\alpha)$ are respectively the first and second derivatives of the gamma function [Eq. (4.24)], which must be evaluated numerically. Equation (4.60) would be implemented by starting with initial guesses for the parameters α and β, perhaps using the moments estimators [Eqs. (4.26)]. Updated values, α^* and β^*, would then result from an application of Eq. (4.60). The updated values would then be substituted into the right-hand side of Eq. (4.60), and the process is repeated until convergence of the algorithm. Convergence would be diagnosed by the parameter estimates changing sufficiently little, perhaps by a small fraction of a percent, between iterations. ¤

Exercises

4.1. Using the binomial distribution as a model for the freezing of Cayuga Lake as presented in Examples 4.1 and 4.2, calculate the probability that the lake will freeze at least once during the 4-year stay of a typical Cornell undergraduate in Ithaca.

4.2. Compute probabilities that one would have to wait
(a) Exactly 5 years for Cayuga Lake to freeze.
(b) At least 25 years for Cayuga Lake to freeze.

4.3. In an article published in the journal *Science*, Gray (1990) contrasts various aspects of Atlantic hurricanes occurring in drought versus wet years in sub-Saharan Africa. During the 18-year drought period 1970–1987, only one "strong" hurricane (intensity 3 or higher) made landfall on the east coast of the United States, but 13 such storms hit the eastern United States during the 23-year wet period 1947–1969.
(a) Assume that the number of hurricanes making landfall in the eastern United States follows a Poisson distribution whose characteristics depend on African rainfall. Fit two Poisson distributions to Gray's data

(one conditional on drought, and one conditional on a wet year, in western Africa).

(b) Compute the probability that at least one strong hurricane will hit the eastern United States, given a dry year in western Africa.

(c) Compute the probability that at least one strong hurricane will hit the eastern United States, given a wet year in western Africa.

4.4. Assume that a "strong" hurricane making landfall in the eastern United States causes, on average, $5 billion in damage. What are the expected values of annual hurricane damage from such storms, according to each of the two conditional distributions in Exercise 4.3?

4.5. Using the June temperature data for Guayaquil, Ecuador, in Table A.3:

(a) Fit a Gaussian distribution.

(b) Without converting the individual data values, determine the two Gaussian parameters that would have resulted if these data had been expressed in degrees Fahrenheit.

(c) Construct a histogram of this temperature data, and superimpose the density function of the fitted distribution on the histogram plot.

4.6. Using the Gaussian distribution and Fig. 4.5:

(a) Estimate the probability that January temperature for Miami (located at the southern tip of Florida) will be colder that 15° C.

(b) What temperature will be higher than all but the warmest 1% of Januarys at Miami?

4.7. For the Ithaca July rainfall data given in Table 4.5:

(a) Fit a gamma distribution using Thom's approximation to the MLEs.

(b) Without converting the individual data values, determine the values of the two parameters that would have resulted if the data had been expressed in millimeters.

(c) Construct a histogram of this precipitation data and superimpose the fitted gamma density function.

Table 4.5

July Precipitation at Ithaca, New York, 1951–1980 (inches)

1951	4.17	1961	4.24	1971	4.25
1952	5.61	1962	1.18	1972	3.66
1953	3.88	1963	3.17	1973	2.12
1954	1.55	1964	4.72	1974	1.24
1955	2.30	1965	2.17	1975	3.64
1956	5.58	1966	2.17	1976	8.44
1957	5.58	1967	3.94	1977	5.20
1958	5.14	1968	0.95	1978	2.33
1959	4.52	1969	1.48	1979	2.18
1960	1.53	1970	5.68	1980	3.43

4.8. Use the result from Exercise 4.7 to compute:
   (a) The 30th and 70th percentiles of July precipitation at Ithaca.
   (b) The difference between the sample mean and the median of the fitted distribution.
   (c) The probability that Ithaca precipitation during any future July will be at least 7 in.

4.9. Using the lognormal distribution to represent the data in Table 4.5, recalculate Exercise 4.8.

4.10. The average of the greatest snow depths for each winter at a location of interest is 80 cm, and the standard deviation (reflecting year-to-year differences in maximum snow depth) is 45 cm.
   (a) Fit a Gumbel distribution to represent these data, using the method of moments.
   (b) Derive the quantile function for the Gumbel distribution, and use it to estimate the snow depth that will be exceeded in only one year out of 100.

4.11. Consider the bivariate normal distribution as a model for the Canandaigua maximum and Canandaigua minimum temperature data in Table A.1.
   (a) Fit the distribution parameters.
   (b) Using the fitted distribution, compute the probability that the maximum temperature will be as cold as or colder than 20° F, given that the minimum temperature is 0° F.

4.12. Construct a Q–Q plot for the temperature data in Table A.3, assuming a Gaussian distribution.

Chapter 5 Hypothesis Testing

5.1 Background

Formal testing of hypotheses, also known as *significance testing*, is generally covered extensively in introductory courses in statistics. Accordingly, this chapter will review only the basic concepts behind formal hypothesis tests, and subsequently emphasize aspects of hypothesis testing that are particularly relevant to applications in the atmospheric sciences.

5.1.1 Parametric versus Nonparametric Tests

There are two contexts in which hypothesis tests are performed; that is, broadly, there are two types of tests. Parametric tests are those conducted in situations where one knows or assumes that a particular theoretical distribution is an appropriate representation for the data and/or the test statistic. The term recognizes that the adjective *parametric* is a more formal synonym for *theoretical*, with regard to probability distributions. Nonparametric tests are conducted without assumptions that particular theoretical forms are appropriate in a given situation.

Very often, parametric tests consist essentially of making inferences about particular distribution parameters. Chapter 4 presented a number of theoretical distributions that have been found to be useful for describing atmospheric data. Fitting such a distribution amounts to distilling the information contained in a sample of data, so that the distribution parameters can be regarded as representing (at least some aspects) of the nature of the underlying physical process of interest. Thus a statistical test concerning a physical process of interest can reduce to a test pertaining to a distribution parameter, such as the Gaussian mean μ.

Nonparametric, or *distribution-free*, tests are constructed without the necessity of assumptions about what, if any, theoretical distribution pertains to the data at hand. Nonparametric tests proceed along one of two basic lines. One approach is to construct the test in such a way that the distribution of the data is unimportant, so that data from any distribution can be treated in the same way. In the following, this approach is referred to as *classical* nonparametric testing, since the methods were devised before the advent of cheap computing power. In the second ap-

proach, crucial aspects of the relevant distribution are inferred directly from the data, by repeated computer manipulations of the observations. These nonparametric tests are known broadly as *resampling* procedures.

5.1.2 The Sampling Distribution

The concept of the sampling distribution is fundamental to all statistical tests. Recall that a statistic is some function computed from a batch of data. A sampling distribution for a statistic is the probability distribution describing batch-to-batch variations of that statistic. Since the batch of data from which any sample statistic (including the test statistic for a hypothesis test) has been computed is subject to sampling variations, sample statistics are subject to sampling variations as well. The value of a statistic computed from a particular batch of data will in general be different from that for the same statistic computed using a different batch of the same kind of data. For example, the average January temperature is obtained by averaging daily temperatures at a particular location for a given year. This statistic is different from year to year.

The random variations of sample statistics can be described using probability distributions just as the random variations of the underlying data can be described using probability distributions. Thus, sample statistics can be viewed as having been drawn from probability distributions. These distributions are called sampling distributions. The sampling distribution provides a probability model describing the relative frequencies of possible values of the statistic.

5.1.3 The Elements of Any Hypothesis Test

Any hypothesis test proceeds according to the following five steps:

1. Identify a *test statistic* that is appropriate to the data and question at hand. The test statistic is the quantity computed from the data values that will be the subject of the test. In parametric settings the test statistic will often be the sample estimate of a parameter of a relevant theoretical distribution. In nonparametric resampling tests there is nearly unlimited freedom in the definition of the test statistic.

2. Define a *null hypothesis*, usually denoted H_0. The null hypothesis constitutes a specific logical frame of reference against which to judge the observed test statistic. Often the null hypothesis will be a "straw man" that one hopes to reject.

3. Define an *alternative hypothesis*, H_A. Many times the alternative hypothesis will be as simple as "H_0 is not true," although more complex alternative hypotheses are also possible.

4. Obtain the *null distribution*, which is simply the sampling distribution of the test statistic given that the null hypothesis is true. Depending on the situation, the null distribution may be an exactly known parametric distribution, a distribution

that is well approximated by a known parametric distribution, or an empirical distribution obtained by resampling the data. Identifying the null distribution is the crucial step defining the hypothesis test.

5. Compare the observed test statistic to the null distribution. If the test statistic falls in a sufficiently improbable region of the null distribution, H_0 is rejected as too unlikely to have been true given the observed evidence. If the test statistic falls within the range of "ordinary" values described by the null distribution, the test statistic is seen as consistent with H_0, which is then not rejected. Note that not rejecting H_0 does not mean that the null hypothesis is true, only that there is insufficient evidence to reject this hypothesis. When H_0 is not rejected one can really only say that it is "not inconsistent" with the observed data.

5.1.4 Test Levels and p Values

The "sufficiently improbable" region of the null distribution just referred to is determined by the *rejection level*, or simply the *level*, of the test. The null hypothesis is rejected if the probability (as represented by the null distribution) of the observed test statistic, *and all other results at least as unfavorable to the null hypothesis*, is less than or equal to the test level. The test level is chosen in advance of the computations, but it depends on the particular investigator's judgment and taste, so that there is usually a degree of arbitrariness about its specific value. Commonly the 5% level is chosen, although tests conducted at the 10% level or the 1% level are not unusual. In situations where penalties can be associated quantitatively with particular test errors (e.g., erroneously rejecting H_0), however, the test level can be optimized (see Winkler, 1972).

The *p value* is the specific probability that the observed value of the test statistic, together with all others at least as unfavorable to the null hypothesis, will occur according to the null distribution. Thus, the null hypothesis is rejected if the p value is less than or equal to the test level, and is not rejected otherwise.

5.1.5 Error Types and the Power of a Test

Another way of looking at the level of a test is as the probability of falsely rejecting the null hypothesis, given that it is true. This false rejection is called a *type I error*, and its probability (the level of the test) is often denoted α. Type I errors are defined in contrast to *type II errors*, which occur if H_0 is not rejected when it is in fact false. The probability of a type II error is usually denoted β.

Figure 5.1 illustrates the relationship between type I and type II errors. A test statistic falling to the right of the rejection level results in rejection of the null hypothesis. Since the area under the probability density function (PDF) of the null distribution to the right of the rejection level (horizontal shading) is 0.05, this is the probability of a type I error. This area is sometimes called the *rejection region*, or the *critical region*. Outcomes in this range are not impossible under H_0, but

Fig. 5.1 Relationship between the rejection level, corresponding to the probability of a type I error, α (horizontal hatching), and the probability of a type II error, β (vertical hatching), for a test conducted at the 5% level. The horizontal axis represents values of the test statistic. Decreasing the probability of a type I error necessarily increases the probability of a type II error, and vice versa.

rather have some small probability α of occurring. Usually the rejection region is defined by that value of the specified probability of a type I error, α, according to the null hypothesis. It is clear from this illustration that, while one would like to minimize the probabilities of both type I and type II errors, this is not in fact possible. Their probabilities, α and β, can be modified by adjusting the level of the test, but decreasing α in this way necessarily increases β.

The level of the test, α, can be prescribed, but the probability of a type II error, β, usually cannot. This is because the alternative hypothesis is defined more generally than the null hypothesis, and usually consists of the union of many specific alternative hypotheses. Whereas α depends on the null distribution, which must be known in order to conduct a test, β depends on which specific alternative hypothesis is actually true, and this is seldom known.

It is sometimes useful, however, to examine the the behavior of β over a range of the alternatives for H_A. This investigation is usually done in terms of the quantity $1 - \beta$, which is known as the *power* of the test against a specific alternative. Geometrically, the power of the test illustrated in Fig. 5.1 is the area under the sampling distribution on the right (for a particular H_A) that does not have vertical shading. The relationship between the power of a test and a continuum of specific alternative hypotheses is called the *power function*. One reason why one might like to choose a less stringent test level (say, $\alpha = 0.10$) would be to better balance error probabilities for a test known to have low power.

5.1.6 One-Sided versus Two-Sided Tests

A statistical test can be either one-sided or two-sided. This dichotomy is sometimes expressed as tests being either one-tailed or two-tailed, since it is the probability in the extremes (tails) of the null distribution that governs whether a test

result is interpreted as being significant. Whether a test is one- or two-sided depends on the nature of the hypothesis being tested.

A one-sided test is appropriate if there is a prior (e.g., a physically based) reason to expect that violations of the null hypothesis will lead to values of the test statistic on a particular side of the null distribution. This situation is illustrated in Fig. 5.1, which has been drawn to imply that alternative hypotheses producing smaller values of the test statistic have been ruled out on the basis of prior information. Here the alternative hypothesis would be stated in terms of the true value being larger than the null hypothesis value (e.g., H_A: $\mu > \mu_0$), rather than the more vague alternative hypothesis that the true value is not equal to the null value (H_A: $\mu \neq \mu_0$). In Fig. 5.1, any test statistic larger than the $100(1 - \alpha)$ percentage point of the null distribution results in the rejection of H_0 at the α level, while very small values of the test statistic do not lead to a rejection of H_0.

A one-sided test is also appropriate when only values on one tail or the other of the null distribution are unfavorable to H_0, because of the way the test statistic has been constructed. For example, a test statistic involving a squared difference will be near zero if the difference is small, but will take on large positive values if the difference is large. In this case, results on the left tail of the null distribution could be quite supportive, indeed, of H_0, so that only right-tail probabilities would be of interest.

Two-sided tests are appropriate when either very large or very small values of the test statistic are unfavorable to the null hypothesis. Usually such tests pertain to the very general alternative hypothesis "H_0 is not true." The rejection region of the null distribution for two-sided tests consists of both the extreme left and extreme right tails of the null distribution. These two portions of the rejection region are delineated in such a way that the sum of their two probabilities under the null distribution yield the level of the test, α. That is, the null hypothesis is rejected at the α level if the test statistic is larger than $100(1 - \alpha)/2\%$ of the null distribution on the right tail, or is smaller than $100(\alpha/2)\%$ of this distribution on the left tail. Thus, a test statistic must be further out on the tail (i.e., more unusual with respect to H_0) to be declared significant in a two-tailed test as compared to a one-tailed test, at a specified test level.

5.1.7 Confidence Intervals: Inverting Hypothesis Tests

Hypothesis testing ideas can be used to construct *confidence intervals* around sample statistics. These are intervals constructed to be wide enough to contain, with a specified probability, the population quantity (often a distribution parameter) corresponding to the sample statistic. In the terminology of Bayesian statistics, these are called *credible intervals*. A typical use of confidence intervals is to construct *error bars* around plotted sample statistics in a graphical display.

In essence, a confidence interval is derived from the hypothesis test whose null hypothesis is that the value of the observed sample statistic corresponds exactly

to the population value it estimates. The confidence interval around this sample statistic then consists of other possible values of the sample statistic for which H_0 would not be rejected. Thus, while hypothesis tests evaluate probabilities associated with an observed test statistic in light of a null distribution, confidence intervals are constructed by finding the values of the test statistic that would not fall into the rejection region. In this sense, confidence interval construction is the inverse operation to hypothesis testing.

Example 5.1. A Hypothesis Test Involving the Binomial Distribution

The testing procedure can be illustrated with a simple, although artificial, example. Suppose that advertisements for a tourist resort in the sunny desert southwest claim that, on average, 6 days out of 7 are cloudless during late summer. To verify this claim, one would need to observe the sky conditions in the area on a number of late summer days, and then compare the fraction observed to be cloudless with the claimed proportion of $6/7 = 0.857$. Assume that one could arrange to take observations on 25 independent occasions. (These will not be consecutive days, because of the serial correlation of daily weather values.) If cloudless skies are observed on 15 of those 25 days, is this observation consistent with, or does it refute, the claim?

This problem fits neatly into the parametric setting of the binomial distribution. A given day is either cloudless or it is not, and observations have been taken sufficiently far apart in time that they can be considered to be independent. By confining observations to only a relatively small portion of the year, one can expect that the probability, p, of a cloudless day is approximately constant from observation to observation.

The first of the five hypothesis testing steps has already been completed, since the test statistic of $x = 15$ out of $N = 25$ days has been dictated by the form of the problem. The null hypothesis is that the resort advertisement was correct in claiming $p = 0.857$. Understanding the nature of advertising, it is reasonable to anticipate that, should the claim be false, the true probability will be lower. Thus the alternative hypothesis is that $p < 0.857$. That is, the test will be "one-tailed," since results indicating $p > 0.857$ are not of interest with respect to discerning the truth of the claim.

Now the crux of the problem is to find the null distribution, that is, the sampling distribution of the test statistic X if the true probability of cloudless conditions is 0.857. This X can be thought of as the sum of 25 independent 0s and 1s, with the 1s having some constant probability of occurring on each of the 25 occasions. These are again the conditions for the binomial distribution. Thus, for this test the null distribution is also binomial, with parameters $p = 0.857$ and $N = 25$.

It remains to compute the probability that 15 or fewer cloudless days would have been observed on 25 independent occasions if the true probability p were in fact 0.857. (This probability is the p-value for the test, which is a different usage for this symbol than the binomial distribution parameter, p). The direct, but tedi-

ous, approach to this computation is summation of the terms given by Eq. (4.1):

$$\Pr\{X \leq 15\} = \sum_{x=0}^{15} \binom{25}{x} 0.857^x (1 - 0.857)^{25-x}. \tag{5.1}$$

Here the terms for the outcomes for $X < 15$ must be included, since observing, say, only 10 cloudless days out of 25 would be even more unfavorable to H_0 than $x = 15$.

A much easier approach to the p-value computation is to use the Gaussian approximation to the binomial distribution. This approximation follows from the central limit theorem, since, as the sum of some number of 0s and 1s, the variable X will follow approximately the Gaussian distribution if N is sufficiently large. Here "sufficiently large" means roughly that $0 < p \pm 2[p(1 - p)/N]^{1/2} < 1$, in which case the binomial X can be characterized approximately using a Gaussian distribution with

$$\mu \approx Np \tag{5.2a}$$

and

$$\sigma \approx \sqrt{Np(1 - p)}. \tag{5.2b}$$

In the current example these are $\mu \approx (25)(0.857) = 21.4$ and $\sigma \approx [(25)(0.857)(1 - 0.857)]^{1/2} = 1.75$, and the relevant standardized Gaussian variable [Eq. (4.20)] is $z = (15 - 21.4)/1.75 = -3.67$. Outcomes more unfavorable to H_0 would result in even more negative values of z, so the p value corresponds to the area under the standard Gaussian density left of -3.67, or $\Phi(-3.67) = 0.00012$. Thus, $x = 15$ is a highly improbable result if the true probability of a cloudless day is 6/7, and this null hypothesis would be resoundingly rejected. According to this test, the observed data are very convincing evidence that the true probability is smaller than 6/7.

For this example, $p + 2[p(1 - p)/N]^{1/2} = 0.997$, which is just barely less than one. Thus the Gaussian approximation may not be as accurate as one would like. The p value for this test as computed from Eq. (5.1) is only 0.0015, which would also lead to rejection of H_0, but indicates that the Gaussian approximation has overstated the significance of this test.

The Gaussian approximation to the binomial, Eqs. (5.2), can also be used to construct a confidence interval (error bars) around the observed estimate of the binomial $p = 15/25 = 0.6$. To do this, one imagines a test whose null hypothesis is that the true binomial probability for this situation is 0.6. This test is then solved in an inverse sense to find the values of the test statistic defining the boundaries of the rejection regions. That is, how large or small a value of x/N would be tolerated before this new null hypothesis would be rejected?

If a 95% confidence region is desired, the test to be inverted will be at the 5% level. Since the true binomial p could be either larger or smaller than the observed x/N, a two-tailed test (rejection regions for both very large and very small x/N)

is appropriate. Referring to Table B.1 since this null distribution is approximately Gaussian, the standardized Gaussian variable cutting off probability equal to $0.05/2 = 0.025$ at the upper and lower tails is $z = \pm 1.96$. (This is the basis of the useful rule of thumb that a 95% confidence interval consists approximately of the mean value ± 2 standard deviations.) Using Eq. (5.2a), the mean number of cloudless days should be $(25)(0.6) = 15$, and from Eq. (5.2b) the corresponding standard deviation is $[(25)(0.6)(1 - 0.6)]^{1/2} = 2.45$. Using Eq. (4.22) with $z = \pm 1.96$ yields $x = 10.2$ and $x = 19.8$, leading to the 95% confidence interval bounded by $p = x/N = 0.408$ and 0.792. Notice that the claimed binomial p of $6/7 = 0.857$ falls outside this interval.

For the test used to construct this confidence interval, $p \pm 2[p(1 - p)/N]^{1/2}$ ranges from 0.404 to 0.796, which is comfortably within the range [0, 1]. The confidence interval computed exactly from the binomial probabilities is [0.40, 0.76], with which the Gaussian approximation agrees very nicely. ¤

5.2 Some Parametric Tests

By far, most of the work on parametric tests in classical statistics has been done in relation to the Gaussian distribution. Tests based on the Gaussian are so pervasive because of the strength of the central limit theorem. As a consequence of this theorem, many non-Gaussian problems can be treated at least approximately in the Gaussian framework. The example test for the binomial parameter p in Example 5.1 is one such case.

5.2.1 One-Sample t Test

Probably the most familiar statistical test is the one-sample t test, which examines the null hypothesis that an observed sample mean has been drawn from a population characterized by some previously specified mean, μ_0. If the number of data values making up the sample mean is sufficiently large that its sampling distribution is essentially Gaussian (by the central limit theorem), then the test statistic

$$t = \frac{\bar{x} - \mu_0}{[\hat{\mathrm{Var}}(\bar{x})]^{1/2}} \tag{5.3}$$

follows a distribution known as *Student's t*, or simply the *t distribution*. Equation (5.3) resembles the standard Gaussian variable Z [Eq. (4.19)], except that a sample estimate of the variance of the sample mean (denoted by the "hat" accent) has been substituted in the denominator.

The t distribution is a symmetrical distribution that is very similar to the standard Gaussian distribution, although with more probability assigned to more extreme values. That is, the t distribution has heavier tails than the Gaussian distribution. The t distribution is controlled by a single parameter, ν, called the *degrees*

of freedom. The parameter ν can take on any positive integer value, with the largest differences from the Gaussian being produced for small values of ν. For the test statistic in Eq. (5.3), $\nu = n - 1$, where n is the number of observations being averaged in the sample mean in the numerator.

Tables of t-distribution probabilities are available in almost any introductory statistics textbook. However, for even moderately large values of ν the t distribution is closely approximated by the standard Gaussian distribution (the differences are about 3% and 1% for $\nu = 30$ and 100, respectively), so it is usually quite acceptable to evaluate probabilities associated with the test statistic in Eq. (5.3) using the standard Gaussian probabilities.

The estimated sample variance of the mean of n independent observations in the denominator of Eq. (5.3) is calculated according to

$$\hat{\mathrm{Var}}[\bar{x}] = \frac{s^2}{n}, \tag{5.4}$$

where s^2 is the sample variance [the square of Eq. (3.5)] of the individual x values being averaged. Equation (5.4) is clearly valid for the simple case of $n = 1$, but also makes intuitive sense for larger values of n. One expects that averaging together, say, pairs ($n = 2$) of x values will give quite irregular results from batch to batch. That is, the sampling distribution of the average of two numbers will have a high variance. Conversely, averaging together batches of $n = 1000$ x values will give very consistent results from batch to batch, because the occasional very large x will tend to be balanced by the occasional very small x: a sample of $n = 1000$ will tend to have nearly equally many very large and very small values. The variance of the sampling distribution of the average of 1000 numbers will thus be small.

For small values of t in Eq. (5.3), the difference in the numerator is small in comparison to the standard deviation of the sampling distribution of the difference, implying a quite ordinary sampling fluctuation that should not trigger rejection of H_0. If the difference in the numerator is more than about twice as large as the denominator in an absolute value sense, the null hypothesis would usually be rejected, corresponding to a two-sided test at the 5% level (cf. Table B.1).

5.2.2 Tests for Differences of Mean under Independence

One of the most common statistical tests is that for the significance of the difference between two independent sample means. Plausible atmospheric examples of this situation might be differences of average winter 500-mb heights when one or the other of two synoptic regimes had prevailed, or perhaps differences in average July temperature at a location as represented in a climate model under a doubling versus no doubling of atmospheric carbon dioxide (CO_2) concentrations.

In general, two sample means consisting of different data drawn from the same population or generating process will be different. The usual test statistic in this situation is a function of the difference of the two sample means being compared,

and the actual observed difference will almost always be some number other than zero. The null hypothesis is usually that the true difference is zero. The alternative hypothesis is either that the true difference is not zero (the case where no a priori information is available as to which underlying mean should be larger, leading to a two-tailed test), or that one of the two underlying means is larger than the other (leading to a one-tailed test). The problem is to find the sampling distribution of the difference of the two sample means, given the null hypothesis assumption that their population counterparts are the same. It is in this context that the observed difference of means can be evaluated for unusualness.

Nearly always—and sometimes quite uncritically—the assumption is tacitly made that the sampling distributions of the two sample means being differenced are Gaussian. This assumption will be true either if the data constituting each of the sample means are Gaussian, or if the sample sizes are sufficiently large that the central limit theorem can be invoked. If both of the two sample means are Gaussian, their difference will be Gaussian as well, since any linear combination of Gaussian variables will itself follow a Gaussian distribution. Under these conditions the test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - E[\bar{x}_1 - \bar{x}_2]}{(\hat{\mathrm{Var}}[\bar{x}_1 - \bar{x}_2])^{1/2}} \tag{5.5}$$

will be distributed as standard Gaussian [Eq. (4.18)] for large samples. Note that this equation has a form similar to both Eqs. (5.3) and (4.20).

If the null hypothesis is equality of means of the two populations from which values of X_1 and X_2 are drawn, then

$$E[\bar{x}_1 - \bar{x}_2] = E[\bar{x}_1] - E[\bar{x}_2] = \mu_1 - \mu_2 = 0. \tag{5.6}$$

Thus, a specific hypothesis about the magnitude of the two means is not required. If some other null hypothesis is appropriate to the problem at hand, that difference of underlying means would be substituted in the numerator of Eq. (5.5).

The variance of a difference (or sum) of two independent random quantities is the sum of the variances of those quantities. Intuitively this makes sense since contributions to the variability of the difference are made by the variability of each of the two quantities being differenced. Thus, with reference to the denominator of Eq. (5.5),

$$\hat{\mathrm{Var}}[\bar{x}_1 - \bar{x}_2] = \hat{\mathrm{Var}}[\bar{x}_1] + \hat{\mathrm{Var}}[\bar{x}_2] = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}, \tag{5.7}$$

where the last equality is achieved using Eq. (5.4). Thus, if the batches making up the two averages are independent, Eq. (5.5) can be transformed to the standard Gaussian Z by rewriting this test statistic as

$$z = \frac{\bar{x}_1 - \bar{x}_2}{[(s_1^2/n_1) + (s_2^2/n_2)]^{1/2}}, \tag{5.8}$$

when the null hypothesis is that the two underlying means μ_1 and μ_2 are equal.

Equation (5.8) is an expression for the familiar two-sample t test. Sometimes a more elaborate version of this test statistic is used, where a different estimate for the variance of the difference of the means is used in the denominator, under the assumption that the corresponding two population variances are equal as well. Under this assumption of equal population variances, Eq. (5.8) becomes instead

$$z = \frac{\bar{x}_1 - \bar{x}_2}{([1/n_1) + (1/n_2)]\{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)\})^{1/2}}, \tag{5.9}$$

where the quantity in curly brackets in the denominator is known as the "pooled" estimate of the population variance. Equation (5.9) follows the t distribution with the degrees-of-freedom parameter $\nu = n_1 + n_2 - 2$, whereas the degrees-of-freedom parameter associated with Eq. (5.8) is somewhat smaller. However, as mentioned previously, it is usually quite acceptable to evaluate probabilities associated with the test statistic in Eq. (5.8) or (5.9) using the standard Gaussian probabilities. For small values of z in Eq. (5.8) or (5.9), the difference of sample means in the numerator is small in comparison to the standard deviation of the sampling distribution of their difference, indicating a quite ordinary value in terms of the null distribution. As before, if the difference in the numerator is more than about twice as large as the denominator in an absolute value sense, the null hypothesis would be rejected at the 5% level for a two-sided test.

It may happen that the elements making up the two averages are *paired*, or observed simultaneously. For example, the daily temperature data in Table A.1 of Appendix A is of this type, since there is an observation of each variable at both locations on each day. When paired atmospheric data are available, the two averages being differenced are generally correlated. Eq. (5.7) [and therefore also Eq. (5.8)] must be modified to account for this correlation. Intuitively, the sampling distribution of the difference in the numerator of Eq. (5.8) will be affected if individual X values going into the averages are strongly correlated. For example, the appropriate cell in Fig. 3.18 indicates that the maximum temperatures at Ithaca and Canandaigua are strongly correlated, so that a relatively warm average monthly maximum temperature at one location would likely be associated with a relatively warm average at the other.

For paired samples the appropriate modification to Eq. (5.7) involves the covariance between the two variables in the denominator of the test statistic:

$$\begin{aligned} \hat{\mathrm{Var}}[\bar{x}_1 - \bar{x}_2] &= \hat{\mathrm{Var}}[\bar{x}_1] + \hat{\mathrm{Var}}[\bar{x}_2] - 2\,\hat{\mathrm{Cov}}[\bar{x}_1, \bar{x}_2] \\ &= \frac{s_1^2}{n} + \frac{s_2^2}{n} - 2\rho_{1,2}\left(\frac{s_1^2}{n}\right)^{1/2}\left(\frac{s_2^2}{n}\right)^{1/2}. \end{aligned} \tag{5.10}$$

Here $\rho_{1,2}$ is simply the Pearson correlation between the x_1 and x_2. Equation (5.7) can be viewed as a special case of Eq. (5.10), since for independent batches of x_1

and x_2, $\rho_{1,2} = 0$. Since $n = n_1 = n_2$ for paired samples, Eq. (5.10) leads to

$$z = \frac{\bar{x}_1 - \bar{x}_2}{[(s_1^2 + s_2^2 - 2\rho_{1,2}s_1s_2)/n]^{1/2}}, \tag{5.11}$$

rather than Eq. (5.8). Typically correlations among atmospheric variables are positive, and Eq. (5.11) is an instance where the positive correlation is beneficial, in that a more sensitive test can be conducted. Here a positive correlation results in a smaller standard deviation for the sampling distribution of the difference of means being tested, implying less underlying uncertainty. This sharper null distribution produces a more powerful test, and allows smaller differences in the numerator to be detected as significantly different from zero.

Intuitively this effect on the sampling distribution of the difference of sample means makes sense as well. Consider again the example of Ithaca and Canandaigua temperatures for January 1987. The positive correlation between daily temperatures at the two locations will result in the batch-to-batch (i.e., January-to-January, or interannual) variations in the two monthly averages moving together for the two locations: Months when Ithaca is warmer than usual tend also to be months when Canandaigua is warmer than usual. The more strongly correlated are x_1 and x_2, the less likely are the pair of corresponding averages from a particular batch of data to differ because of sampling variations. To the extent that the two sample averages are different, then, the evidence against their underlying means not being the same is stronger, as compared to the situation when $\rho_{1,2}$ is near zero.

5.2.3 Tests for Differences of Mean under Serial Dependence

The material in the previous sections is essentially a recapitulation of the classical tests for comparing sample means, presented in almost every elementary statistics textbook. A key assumption in the foregoing is the independence among the individual observations making up each of the sample means in the test statistic. That is, it is assumed that all the x_1 values are mutually independent and that the x_2 values are mutually independent. This assumption of independence leads to the expression in Eq. (5.4) that allows computation of the variance of the null distribution.

Unfortunately, atmospheric data often do not satisfy the independence assumption. Most frequently the averages to be tested are time averages, and the persistence, or time dependence, often exhibited is the cause of the violation of the assumption of independence. Lack of independence invalidates Eq. (5.4). In particular, *meteorological persistence* implies that the variance of a time average is larger than specified by Eq. (5.4). Ignoring the time dependence thus leads to an underestimate of the variance of sampling distribution of the test statistic in Eq. (5.3), (5.5), or (5.11). This underestimate leads in turn to an inflated value of the test statistic, and consequently to overconfidence regarding the significance

Fig. 5.2 Comparison of artificial time series of 100 independent Gaussian values (a) with 100 autocorrelated Gaussian values having $\rho_1 = 0.6$ (b). Both series were drawn from a generating process with $\mu = 0$. Nearby values in the autocorrelated series tend to be more alike, with the result that averages over subsections of the autocorrelated time series are more likely to be far from zero than are averages from the independent series. This implies that sampling distributions of averages computed from the autocorrelated series will have larger variances.

of the difference in the numerator. Equivalently, properly representing the effect of persistence in the data will require larger sample sizes to reject the null hypothesis for a given magnitude of the difference in the numerator.

Figure 5.2 may help to understand why serial correlation leads to a larger variance for the sampling distribution of a time average. Figure 5.2a is an artificial time series of 100 independent Gaussian variables drawn from a generating process with $\mu = 0$. The series in Fig. 5.2b also consists of Gaussian variables having $\mu = 0$, but in addition this series has a lag-1 autocorrelation [Eq. (3.23)] of $\rho_1 = 0.6$. This value of the autocorrelation was chosen here because it is typical of the autocorrelation exhibited by daily temperatures (e.g., Madden, 1979). The two plots appear similar because the autocorrelated series was generated from the independent series according to what is called a first-order autoregressive process (see Chapter 8).

The outstanding difference between the independent and autocorrelated pseudodata in Fig. 5.2 is that the correlated series is smoother, so that adjacent and nearby values tend to be more alike than in the independent series. The autocorrelated series exhibits longer runs of points away from the (population) mean value. As a consequence, averages computed over subsets of the autocorrelated record are less likely to contain compensating points with large absolute value but of different sign, and those averages are therefore more likely to be far from zero

(the true underlying average) than their counterparts computed using the independent values. That is, these averages will be less consistent from batch to batch. This is just another way of saying that the sampling distribution of an average of autocorrelated data has a higher variance than that of independent data.

Not surprisingly, the problem of estimating the variance of the sampling distribution of time averages has received considerable attention in the meteorological literature (e.g., Jones, 1975; Katz, 1982; Madden, 1979; Zwiers and Thiébaux, 1987). One convenient and practical approach to dealing with the problem is to think in terms of the *effective sample size*, or *equivalent number of independent samples*, n'. That is, imagine that there is a fictitious sample size, $n' < n$ of independent values, for which the sampling distribution of the average has the same variance as the sampling distribution of the average over the n autocorrelated values at hand. Then, n' could be substituted for n in Eq. (5.4), and the classical tests described in the previous section could be carried through as before.

Estimation of the effective sample size is most easily approached if it can be assumed that the underlying data follow a first-order autoregressive process (Jones, 1975; Katz, 1985). It turns out that first-order autoregressions (Chapter 8) are often reasonable approximations for representing the persistence of daily meteorological values. This assertion can be appreciated informally by looking at Fig. 5.2b. This plot consists of random numbers, but resembles statistically the day-to-day fluctuations in a meteorological variable like surface temperature.

The persistence in a first-order autoregression is completely characterized by the single parameter ρ_1, the lag-1 autocorrelation coefficient, which can be estimated from a data series using the sample estimate, r_1 [Eq. (3.23)]. Using this correlation, the effective sample size can be estimated using the approximation

$$n' \cong n\frac{1 - \rho_1}{1 + \rho_1}. \tag{5.12}$$

When there is no time correlation, $\rho_1 = 0$ and $n' = n$. As ρ_1 increases, the effective sample size becomes progressively smaller. When a more complicated time-series model is necessary to describe the persistence, appropriate but more complicated expressions for the effective sample size can be derived (see Katz, 1982, 1985; Chapter 8).

Using Eq. (5.12), the counterpart to Eq. (5.4) for the variance of a time average over a sufficiently large sample becomes

$$\hat{\mathrm{Var}}[\bar{x}] \cong \frac{s^2}{n'} = \frac{s^2}{n}\left(\frac{1 + \rho_1}{1 - \rho_1}\right). \tag{5.13}$$

The ratio $(1 + \rho_1)/(1 - \rho_1)$ acts as a *variance inflation factor*, adjusting the variance of the sampling distribution of the time average to reflect the influence of the serial correlation. Sometimes this variance inflation factor is called the *time between effectively independent samples*, T_0 (e.g., Leith, 1973). Equation (5.4) can be seen as a special case of Eq. (5.13), with $\rho_1 = 0$.

Example 5.2. Two-Sample t Test for Correlated Data

Consider testing whether the average maximum temperatures at Ithaca and Canandaigua for January 1987 (Table A.1 in Appendix A) are significantly different. This is equivalent to testing whether the difference of the two sample means is significantly different from zero, so that Eq. (5.6) will hold for the null hypothesis. It has been shown previously (Fig. 3.5) that these two batches of daily data are reasonably symmetrical and well-behaved, so the sampling distribution of the monthly average should be nearly Gaussian under the central limit theorem. Thus, the parametric test just described (which assumes the Gaussian form for the sampling distribution) should be appropriate.

The data for each location were observed on the same 31 days in January 1987, so the two batches constitute paired samples. Equation (5.10) is therefore the appropriate choice for the variance of the sampling distribution of the test statistic. Furthermore, we know that the daily data underlying the two time averages exhibit serial correlation (Fig. 3.15 for the Ithaca data), so the effective sample size corrections in Eqs. (5.12) and (5.13) will be necessary as well.

Table 5.1 contains the sample statistics necessary to perform the test. The effective sample size, n', is substituted for n in Eq. (5.10), and Eqs. (5.10) and (5.6) are substituted into Eq. (5.5). These substitutions yield the test statistic

$$z = \frac{(29.87 - 31.77) - 0}{[7.71^2/9.82 + 7.86^2/7.49 - (2)(0.957)(7.71^2/9.82)^{1/2}(7.86^2/7.49)^{1/2}]^{1/2}} = -2.15, \tag{5.14}$$

which will follow approximately the standard Gaussian distribution if the null hypothesis is true. From Table B.1, $\Pr\{Z \leq -2.15\} = 0.016$. If one knew in advance (on physical grounds, since the Canandaigua station is at lower elevation and closer to Lake Ontario), that any difference in means would derive from Canandaigua being warmer, the test would be one-sided, the p value would be 0.016, and the null hypothesis of equality of sample means would be rejected at

Table 5.1

Sample Statistics for the January 1987 Maximum Temperature Data from Appendix A Required to Test for Equality of Ithaca and Canandaigua Average Maximum Temperatures Using Eq. (5.5)

	Ithaca (Ith.)		Canandaigua (Can.)
$\bar{x}$	29.87°F		31.77°F
s	7.71°F		7.86°F
r_1	0.519		0.611
n'	9.82		7.49
$r_{\text{Ith.–Can.}}$		0.957	

the 5% level. If, on the other hand, one had no prior information about the physical relationship between the stations, the test would be two-tailed. In this case large positive values of z would be just as damaging to H_0 as large negative ones. For a two-tailed test the p value would be $(2)(0.016) = 0.032$, which would still lead to rejection of H_0 at the 5% level. ¤

5.2.4 Goodness-of-Fit Tests

When discussing the fitting of parametric distributions to data samples in Chapter 4, methods for visually and subjectively assessing the goodness of fit were presented. Formal, quantitative tests of the goodness of fit also exist, and these are carried out within the framework of hypothesis testing. However, the graphical methods can still be useful when formal tests are conducted, for example, in pointing out where and how lack of fit is manifested. Many goodness-of-fit tests have been devised, but only a few common ones are presented here.

A very frequently used test of the goodness of fit of a theoretical distribution is the one-sample Kolmogorov–Smirnov (K-S) test. This test consists of comparing the empirical and theoretical CDFs for a single sample, under the null hypothesis that the observed data were drawn from the theoretical distribution. A sufficiently large discrepancy will result in the null hypothesis being rejected.

In its original form, the K-S test is applicable to any distributional form (i.e., to any of the theoretical distributions presented in Chapter 4), provided that the parameters have *not* been estimated from the data sample. In practice this provision can constitute a serious limitation to the use of the original K-S test, since it is often the correspondence between a fitted distribution and the particular batch of data used to fit it that is of interest. While this may seem like a trivial problem, it can have serious consequences, as has been pointed out by Crutcher (1975). Estimating the distribution parameters from the same batch of data used to test the goodness of fit results in the particular fitted distribution parameters being "tuned" to the data sample. When erroneously using K-S tables that assume independence between the test data and the estimated parameters, it will often be the case that the null hypothesis (that the distribution fits well) will not be rejected when in fact it should be.

With modification, the K-S framework can be used in situations where the distribution parameters have been fit to the same data used in the test. In this situation, the K-S test is often called the *Lilliefors test*, after the statistician who did much of the early work on the subject. Both the original K-S test and the Lilliefors test utilize the test statistic

$$D_n = \max_x |F_n(x) - F(x)|, \tag{5.15}$$

where $F_n(x)$ is the empirical cumulative probability, estimated as $F_n(x_{(i)}) = i/n$ for the ith smallest data value; and $F(x)$ is the theoretical cumulative distribution

Table 5.2

Critical Values for the K-S Statistic D_n to Assess Goodness of Fit of Gamma Distributions, as a Function of the Estimated Shape Parameter, α, When Distribution Parameters Have Been Fit Using the Data to Be Tested[a]

	20% level			10% level			5% level			1% level		
α	$n = 25$	$n = 30$	Large n	$n = 25$	$n = 30$	Large n	$n = 25$	$n = 30$	Large n	$n = 25$	$n = 30$	Large n
1	0.165	0.152	$0.84/\sqrt{n}$	0.185	0.169	$0.95/\sqrt{n}$	0.204	0.184	$1.05/\sqrt{n}$	0.241	0.214	$1.20/\sqrt{n}$
2	0.159	0.146	$0.81/\sqrt{n}$	0.176	0.161	$0.91/\sqrt{n}$	0.190	0.175	$0.97/\sqrt{n}$	0.222	0.203	$1.16/\sqrt{n}$
3	0.148	0.136	$0.77/\sqrt{n}$	0.166	0.151	$0.86/\sqrt{n}$	0.180	0.165	$0.94/\sqrt{n}$	0.214	0.191	$1.80/\sqrt{n}$
4	0.146	0.134	$0.75/\sqrt{n}$	0.164	0.148	$0.83/\sqrt{n}$	0.178	0.163	$0.91/\sqrt{n}$	0.209	0.191	$1.06/\sqrt{n}$
8	0.143	0.131	$0.74/\sqrt{n}$	0.159	0.146	$0.81/\sqrt{n}$	0.173	0.161	$0.89/\sqrt{n}$	0.203	0.187	$1.04/\sqrt{n}$
∞	0.142	0.131	$0.736/\sqrt{n}$	0.158	0.144	$0.805/\sqrt{n}$	0.173	0.161	$0.886/\sqrt{n}$	0.200	0.187	$1.031/\sqrt{n}$

[a] The row labeled "$\alpha = \infty$" pertains to the Gaussian distribution with parameters estimated from the data.

[From Crutcher (1975), reproduced with permission of the American Meteorological Society.]

function (CDF) evaluated at x [Eq. (4.12)]. Thus the K-S test statistic D_n looks for the largest difference, in an absolute value sense, between the empirical and fitted CDFs.

Any real and finite batch of data will exhibit sampling fluctuations resulting in a nonzero value for D_n, even if the null hypothesis is true and the theoretical distribution fits very well indeed. If D_n is sufficiently large, the null hypothesis can be rejected. How large is large enough depends on the level of the test, of course, but also on the sample size and the particular theoretical form being fit. Commonly it is the goodness of fit of a Gaussian distribution that is being assessed, but other distributions may also be of interest. Table 5.2, from Crutcher (1975), lists critical values of D_n (above which the null hypothesis would be rejected) for four test levels for the gamma distribution. These critical values depend on both the sample size and the estimated shape parameter, α. Larger samples will be less subject to irregular sampling variations, so the tabulated critical values decline for larger n. That is, smaller maximum deviations from the the fitted theoretical distribution [Equation (5.15)] are tolerated for larger sample sizes. Critical values in the last row of the table, for "$\alpha = \infty$," pertain to the Gaussian distribution, since as the gamma shape parameter becomes very large the gamma distribution converges toward the Gaussian.

Example 5.3. Comparing Gaussian and Gamma Fits Using the K-S Test

Consider the fits of the gamma and Gaussian distributions to the 1933–1982 Ithaca January precipitation data, from Table A.2, shown in Fig. 4.14. The approximate MLEs for the gamma distribution parameters [Eq. (4.28) or (4.30a), and Eq. (4.34)] are $\alpha = 3.76$ in. and $\beta = 0.52$ in. The sample mean and standard deviation (i.e., the Gaussian parameter estimates) for these data are 1.96 and 1.12 in., respectively. These two fitted PDFs are superimposed on the data histogram in Fig. 4.14. Figure 5.3 illustrates the K-S (actually, the Lilliefors) test for these two fitted distributions. In each panel (a and b) of Fig. 5.3, the dots are the empirical cumulative probability estimates, $F_n(x)$, and the smooth curves are the fitted theoretical cdfs, $F(x)$, both plotted as functions of the observed monthly precipitation. The maximum differences between the empirical and fitted theoretical CDFs occur at the same (highlighted) point, yielding $D_n = 0.068$ for the gamma distribution (a) and $D_n = 0.131$ for the Gaussian distribution (b).

This example requires that two tests be conducted: one each for the goodness of fit of the Gaussian and gamma distributions. In each case the null hypothesis is that the precipitation data were drawn from the fitted distribution, and the alternative hypothesis is that they were not. These will necessarily be one-sided tests, because the test statistic D_n is the absolute value of the largest difference between the theoretical and empirical cumulative probabilities. Therefore values of the test statistic on the far right tail of the null distribution will indicate large discrepancies that are unfavorable to H_0, whereas values of the test statistic on the left tail of the null distribution will indicate $D_n \approx 0$, or near-perfect fits that are very supportive indeed of the null hypothesis.

Fig. 5.3 The Kolmogorov-Smirnov D_n statistic as applied to the 1933–1982 Ithaca January precipitation data (dots), fitted to a gamma distribution (a) and a Gaussian distribution (b). Solid curves indicate theoretical cumulative distribution functions, and dots show the corresponding empirical estimates. The maximum difference between the empirical and theoretical CDFs occurs for the highlighted point, and is substantially greater for the Gaussian distribution.

The critical values in Table 5.2 are the minimum D_n necessary to reject H_0: the leftmost bounds of the relevant rejection, or critical regions. The sample size of $n = 50$ is sufficient to evaluate the tests using critical values from the "large n" columns. In the case of the Gaussian distribution, the relevant row of the table is for $\alpha = \infty$. Since $0.886/\sqrt{50} = 0.125$ and $1.031/\sqrt{50} = 0.146$ bound the observed $D_n = 0.131$, the null hypothesis that the precipitation data were drawn from this Gaussian distribution would be rejected at the 5% level, but not the 1% level. For the fitted gamma distribution the nearest row in Table 5.2 is for $\alpha = 4$, where even at the 20% level the critical value of $0.75/\sqrt{50} = 0.106$ is substantially higher than the observed $D_n = 0.068$. Thus the data are quite consistent with the proposition of their having been drawn from this gamma distribution. ¤

A related test is the two-sample K-S test, or Smirnov test. Here the idea is to compare two batches of data to one another under the null hypothesis that they were drawn from the same (but unspecified) distribution or generating process. The Smirnov test statistic

$$D_S = \max_x |F_n(x_1) - F_m(x_2)|, \tag{5.16}$$

looks for the largest (in absolute value) difference between the empirical CDFs of samples of n observations of X_1 and m observations of X_2. Again, because of the taking of absolute values in Eq. (5.16), the test is one-sided, and the null hypothesis that the two data samples were drawn from the same distribution is rejected at the $\alpha \cdot 100\%$ level if

$$D_S > \left[\frac{1}{2}\left(\frac{1}{n} + \frac{1}{m}\right)\ln\left(\frac{\alpha}{2}\right)\right]^{1/2}. \tag{5.17}$$

Note that in Eq. (5.17), α is the probability of a type I error, (cf. Fig. 5.1), and not a distribution parameter.

Another common goodness-of-fit test is the χ^2 test. While the K-S test can be thought of as comparing empirical and theoretical CDFs, the χ^2 test essentially compares the data histogram with the probability distribution (for discrete variables) or density (for continuous variables) function. It is less sensitive to discrepancies in the extreme tails than is the K-S test. The χ^2 test actually operates more naturally for discrete random variables, since to implement it, the range of the data must be divided into discrete classes, or bins. For continuous random variables, the probability density function is integrated over each of some number of MECE classes to obtain the theoretical probabilities of observations in each class. The test statistic involves the counts of data values falling into each class in relation to the computed theoretical probabilities,

$$\begin{aligned}\chi^2 &= \sum_{\text{classes}} \frac{(\#\text{ observed} - \#\text{ expected})^2}{\#\text{ expected}} \\ &= \sum_{\text{classes}} \frac{(\#\text{ observed} - n\ \Pr\{\text{data in class}\})^2}{n\ \Pr\{\text{data in class}\}}.\end{aligned} \tag{5.18}$$

In each class, the number (#) of data values "expected" to occur, according to the fitted distribution, is simply the probability of occurrence in that class times the sample size, n. The number of expected occurrences need not be an integer value. If the fitted distribution is very close to the data, the expected and observed counts will be very close for each class; and the squared differences in the numerator of Eq. (5.18) will all be very small, yielding a small χ^2. If the fit is not good, at least a few of the classes will exhibit large discrepancies. These will be squared in the numerator of Eq. (5.18) and lead to large values of χ^2. It is not necessary for the classes to be of equal width or equal probability, but classes with small numbers

of expected counts should be avoided. Sometimes a minimum of five "expected" events per class is imposed.

Under the null hypothesis that the data were drawn from the fitted distribution, the test statistic follows the χ^2 distribution with parameter ν = (# of classes − # of parameters fit − 1) degrees of freedom. The test will be one-sided, again because the test statistic is confined to positive values by the squaring process in the numerator of Eq. (5.18), and small values of the test statistic support H_0. The χ^2 distribution is actually a special case of the gamma distribution, with $\alpha = \nu/2$ and $\beta = 2$. Thus, for $\nu \leq 10$ Table B.2 can be used to find χ^2 probabilities.

Example 5.4. Comparing Gaussian and Gamma Distribution Fits Using the χ^2 Test

Again consider the fits of the gamma and Gaussian distributions to the 1933–1982 Ithaca January precipitation data in Table A.2. The two fits are illustrated in relation to the data in Fig. 4.14. Table 5.3 contains the necessary information to conduct the χ^2 test for these two distributions. The precipitation amounts have been divided into six classes, or bins, the limits of which are indicated in the first row of the table. The second row indicates the number of years in which the January precipitation total was within each class. Both the fitted gamma distribution, with $\alpha = 3.76$ in. and $\beta = 0.52$ in.; and the fitted Gaussian distribution, with $\mu = 1.96$ in. and $\sigma = 1.12$ in.; have been integrated over these classes to obtain probabilities for precipitation in each class. These probabilities were then multiplied by $n = 50$ to obtain the "expected" number of counts.

Applying Eq. (5.18) yields $\chi^2 = 5.05$ for the gamma distribution and $\chi^2 = 14.96$ for the Gaussian distribution. As was also evident from the graphical fit in Fig. 4.14 and from Example 5.3, these test statistics indicate that the Gaussian distribution fits this precipitation data substantially less well. Under the respective

Table 5.3

The χ^2 Goodness-of-Fit Test Applied to Gamma and Gaussian Distributions Fit to the 1933–1982 Ithaca January Precipitation Data[a]

Class (inches)	<1	1–1.5	1.5–2	2–2.5	2.5–3	≥3
Observed #	5	16	10	7	7	5
Gamma						
Probability	0.161	0.215	0.210	0.161	0.108	0.145
Expected #	8.05	10.75	10.50	8.05	5.40	7.25
Gaussian						
Probability	0.195	0.146	0.173	0.173	0.132	0.176
Expected #	9.75	7.30	8.65	8.90	6.60	8.80

[a] "Expected" numbers (#) of occurrences in each bin are obtained by multiplying the respective probabilities by $n = 50$.

null hypotheses, these two test statistics are drawn from a χ^2 distribution with degrees of freedom $\nu = 6 - 2 - 1 = 3$. Table 5.3 contains six classes, and two parameters (α and β, or μ and σ, for the gamma and Gaussian, respectively) were fit for each distribution. The null distribution is therefore a gamma distribution with $\alpha = \nu/2 = 1.5$, and $\beta = 2$.

To obtain p values using Table B.2, it is necessary to standardize the test statistics by dividing by the scale parameter β (see Chapter 4), yielding $\xi_{\text{gamma}} = 5.05/2 = 2.53$, and $\xi_{\text{Gauss}} = 14.96/2 = 7.48$. Interpolating on the $\alpha = 1.5$ row in Table B.2, $\xi_{\text{gamma}} = 2.53$ corresponds to a cumulative probability of 0.826, or a p value of $1 - 0.826 = 0.174$. The null hypothesis that the data were drawn from this gamma distribution would again not be rejected. For the fitted Gaussian distribution, $\xi_{\text{Gauss}} = 7.48$ corresponds to a cumulative probability that is substantially higher than 0.99, again providing strong evidence that these data were not drawn from this Gaussian distribution. ¤

5.2.5 Likelihood Ratio Test

Sometimes one needs to construct a test in a parametric setting, but the hypothesis is sufficiently complex that the simple, familiar parametric tests cannot be brought to bear. A flexible alternative, known as the *likelihood ratio test*, can be used if two conditions are satisfied. First, one must be able to cast the problem in such a way that the null hypothesis pertains to some number, k_0, of free (i.e., fitted) parameters, and the alternative hypothesis pertains to some larger number, $k_A > k_0$, of parameters. Second, it must be possible to regard the k_0 parameters of the null hypothesis as a special case of the "full" parameter set of k_A parameters. Examples of this second condition on H_0 could include forcing some of the k_A parameters to have constant values, or imposing equality between two or more of them. As the name implies, the likelihood ratio test compares the likelihoods associated with H_0 versus H_A, when the k_0 and k_A parameters, respectively, have been fit using the method of maximum likelihood (Chapter 4).

Even if the null hypothesis is true, the likelihood associated with H_A will always be at least as large as that for H_0. This is because the greater number of parameters $k_A > k_0$ allows the likelihood function for the former greater freedom in accommodating the observed data. The null hypothesis is therefore rejected only if the likelihood associated with the alternative is sufficiently large that the difference is unlikely to have resulted from sampling variations.

The test statistic for the likelihood ratio test is

$$\Lambda = 2 \ln\left[\frac{\mathfrak{l}(H_A)}{\mathfrak{l}(H_0)}\right] = 2[L(H_A) - L(H_0)]. \tag{5.19}$$

Here $\mathfrak{l}(H_0)$ and $\mathfrak{l}(H_A)$ are the likelihood functions (Section 4.7) associated with

the null and alternative hypotheses, respectively. The second equality, involving the difference of the log-likelihoods $L(H_0)$ and $L(H_A)$, is used in practice since it is generally the log-likelihoods that are maximized (and thus computed) when fitting the parameters.

Under H_0, and given a large sample size, the sampling distribution of Λ is χ^2, with degrees of freedom $\nu = k_A - k_0$. That is, the degrees-of-freedom parameter is given by the difference between H_A and H_0 in the number of empirically estimated parameters. Again, the χ^2 is a special case of the gamma distribution, and the null distribution of Λ will be a gamma distribution with $\alpha = \nu/2 = (k_A - k_0)/2$ and $\beta = 2$. Since small values of Λ are not unfavorable to H_0, the test is one-sided, and H_0 is rejected only if the observed Λ is in a sufficiently improbable region on the right tail.

Example 5.5. Testing for Climate Change Using the Likelihood Ratio Test

Suppose there was reason to suspect that the first 25 years (1933–1957) of the Ithaca January precipitation data in Table A.2 had been drawn from a gamma distribution different from that in the second half (1958–1982). This question can be tested against the null hypothesis that all 50 precipitation totals were drawn from the same gamma distribution, using the likelihood ratio test. To perform the test it is necessary to fit gamma distributions separately to the two halves of the data, and compare these fitted distributions with the single gamma distribution fit using the full data set.

The relevant information is presented in Table 5.4, which indicates some differences between the two 25-year periods. For example, the average January precipitation ($= \alpha\beta$) for 1933–1957 was 1.87 in., whereas the corresponding average for 1958–1982 was 2.05 in. The year-to-year variability ($= \alpha\beta^2$) of January precipitation was greater in the second half of the period as well. Whether the extra two parameters required to represent the January precipitation, using two gamma distributions rather than one, are justified by the data can be evaluated using the test statistic in Eq. (5.19). For this specific problem the test statistic is

Table 5.4

Gamma Distribution Parameters (MLEs) and Log-Likelihoods for Fits to First and Second Halves of the 1933–1982 Ithaca January Precipitation Data and to the Full Data Set

Dates	α	β	$L(\alpha,\beta)$
H_A			
1933–1957	4.525	0.4128	−30.2796
1958–1982	3.271	0.6277	−35.8965
H_O			
1933–1982	3.764	0.5209	−66.7426

$$\Lambda = 2\left\{\left[\sum_{i=1933}^{1957} L_\Gamma(\alpha_1, \beta_1; x_i)\right] + \left[\sum_{j=1958}^{1982} L_\Gamma(\alpha_2, \beta_2; x_j)\right] - \left[\sum_{i=1933}^{1982} L_\Gamma(\alpha_0, \beta_0; x_i)\right]\right\}, \tag{5.20}$$

where the subscripts 1, 2, and 0 on the parameters refer to the first half, the second half, and the full period (null hypothesis), respectively, and the log-likelihood for the gamma distribution given a single observation, x, is [cf. Eq. (4.23)]

$$L_\Gamma(\alpha, \beta; x) = (\alpha - 1) \ln \frac{x}{\beta} - \frac{x}{\beta} - \ln \beta - \ln \Gamma(\alpha). \tag{5.21}$$

It is the three terms in square brackets in Eq. (5.20) that are given in the last column of Table 5.4.

Using the information in Table 5.4, $\Lambda = 2\ (-30.2796 - 35.8965 + 66.7426) = 1.130$. Since there are $k_A = 4$ parameters under $H_A\ (\alpha_1, \beta_1, \alpha_2, \beta_2)$ and $k_0 = 2$ parameters under $H_0\ (\alpha_0, \beta_0)$, the null distribution is the χ^2 distribution with $\nu = 2$. This is a gamma distribution having $\alpha = \nu/2 = 1$ and $\beta = 2$. Interpolating on the $\alpha = 1$ row of Table B.2, one finds that $F(1.130/2) = 0.43$, so that the p value (tail area to the right of the test statistic) is $1 - 0.43 = 0.57$. One concludes that the observed Λ is quite ordinary in the context of the null hypothesis that the two data records were drawn from the same gamma distribution, which would not be rejected. ¤

5.3 Nonparametric Tests

Not all formal hypothesis tests rest on assumptions involving theoretical distributions for the data, or theoretical sampling distributions of the test statistics. Tests not requiring such assumptions are called *nonparametric*, or *distribution-free*. Nonparametric methods are appropriate if either or both of the following conditions apply:

1. One knows or suspects that the parametric assumption(s) required for a particular test are not met, for example grossly non-Gaussian data in conjunction with the t test for the difference of means in Eq. (5.5).
2. A test statistic that is suggested or dictated by the physical problem at hand is a complicated function of the data, and its sampling distribution is unknown and/or cannot be derived analytically.

The fundamental ideas of hypothesis testing apply to both parametric and nonparametric tests. In particular, the five elements of the hypothesis test presented at

the beginning of this chapter hold also for nonparametric tests. The difference between parametric and nonparametric tests is in the means by which the null distribution is obtained in step 4.

One can recognize two branches of nonparametric testing. The first, called *classical* nonparametric testing in the following, consists of tests based on mathematical analysis of selected hypothesis test settings. These are older methods, devised before the advent of cheap and widely available computing. These methods employ analytic mathematical results (formulas) that are applicable to data drawn from any distribution. Only a few classical nonparametric tests for location will be presented here, although the range of classical nonparametric methods is much more extensive (e.g., Daniel, 1990).

The second branch of nonparametric testing is comprised of procedures collectively called *resampling* tests. Resampling tests build up a discrete approximation to the null distribution using a computer, by repeatedly operating on (resampling) the data set at hand. Since the null distribution is arrived at empirically, the analyst is free to use virtually any test statistic that may be relevant, regardless of how mathematically complicated it may be.

5.3.1 "Classical" Nonparametric Tests for Location

Two classical nonparametric tests for the differences in location between two data samples are especially common and useful. These are the Wilcoxon–Mann–Whitney rank-sum test for two independent samples [analogous to the parametric test in Eq. (5.8)], and the Wilcoxon signed-rank test for paired samples [corresponding to the parametric test in Eq. (5.11)].

The Wilcoxon–Mann–Whitney rank-sum test was devised independently in the 1940s by Wilcoxon, and by Mann and Whitney, although in different forms. The notations from both forms of the test are commonly used, and this can be the source of some confusion. However, the fundamental idea behind the test is not difficult to understand. The test is resistant, in the sense that a few "wild" data values that would completely invalidate the t test of Eq. (5.8) will have little or no influence. It is robust in the sense that, even if all the assumptions required for the t test in Eq. (5.8) are met, the rank-sum test is almost as good.

Given two batches of independent (i.e., both serially independent, and unpaired) data, the aim is to test for a possible difference in location. Here *location* is meant in the EDA sense of overall magnitude, or the nonparametric analog of the mean. The null hypothesis is that the two data batches have the same location. Both one-sided (one batch is expected in advance to be larger or smaller than the other) and two-sided (no prior information on which batch should be larger) alternative hypotheses are possible.

Under the null hypothesis that the two data batches are from the same distribution, the labeling of each data value as belonging to one or the other batch is

entirely arbitrary. That is, if the two batches are really subsamples of the same population, each observation is as likely to have been placed in one batch as the other by the sampling process that generated the data. Under the null hypothesis, then, there are not n_1 observations in batch 1 and n_2 observations in batch 2, but rather $n = n_1 + n_2$ observations making up a single empirical distribution.

The rank-sum test statistic is a function not of the data values themselves, but of their ranks within the n observations that are pooled under the null hypothesis. It is this feature that makes the underlying distribution(s) of the data irrelevant. Define R_1 as the sum of the ranks held by the members of batch 1 in this pooled distribution, and R_2 as the sum of the ranks held by the members of batch 2. Since there are n members of the pooled empirical distribution implied by the null distribution, $R_1 + R_2 = 1 + 2 + 3 + 4 + \cdots + n = (n)(n + 1)/2$. If the two batches really do have the same location (i.e., if H_0 is true), then R_1 and R_2 will be similar in magnitude if $n_1 = n_2$. Regardless of whether the sample sizes are equal, however, R_1/n_1 and R_2/n_2 should be similar in magnitude if the null hypothesis is true.

The null distribution for R_1 and R_2 is obtained in a way that exemplifies the approach of nonparametric tests more generally. If the null hypothesis is true, the observed partitioning of the data into two groups of size n_1 and n_2 is only one of very many equally likely ways in which the n values could have been split and labeled. Specifically, there are $(n!)/[(n_1!)(n_2!)]$ such equally likely partitions of the data under the null hypothesis. If $n_1 = n_2 = 10$, this number of possible distinct pairs of samples is 184,756. Conceptually, imagine the statistics R_1 and R_2 being computed for each of these 184,756 possible arrangements of the data. It is simply this very large collection of (R_1, R_2) pairs that constitutes the null distribution. If the observed R_1 and R_2 fall comfortably within this large empirical distribution, then that particular partition of the n observations is quite consistent with H_0. If, however, the observed R_1 and R_2 are more different from each other than under most of the other possible partitions of the data, H_0 would be rejected.

It is not actually necessary to compute the test statistic for all $(n!)/[(n_1!)(n_2!)]$ possible arrangements of the data. Rather, the Mann–Whitney U statistic,

$$U_1 = R_1 - \frac{n_1}{2}(n_1 + 1) \tag{5.22a}$$

or

$$U_2 = R_2 - \frac{n_2}{2}(n_2 + 1), \tag{5.22b}$$

is computed for one or the other of the two Wilcoxon rank-sum statistics, R_1 or R_2. Both U_1 and U_2 carry the same information, since $(U_1 + U_2) = (n_1)(n_2)$, although some tables of null distribution probabilities for the rank-sum test require the smaller of U_1 and U_2. For even moderately large values of n_1 and n_2 (both larger than about 10), however, a simple method for evaluating null distribution

probabilities is available. In this case, the null distribution of the Mann-Whitney U statistic is approximately Gaussian, with

$$\mu_U = \frac{n_1 n_2}{2} \tag{5.23a}$$

and

$$\sigma_U = \left[\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right]^{1/2}. \tag{5.23b}$$

Example 5.6. Evaluation of a Cloud Seeding Experiment Using the Wilcoxon–Mann–Whitney Test

Table 5.5 contains data from a weather modification experiment investigating the effect of cloud seeding on lightning strikes (Baughman *et al.*, 1976). It was suspected in advance that seeding the storms would reduce lightning. The experimental procedure involved randomly seeding or not seeding candidate thunderstorms, and recording a number of characteristics of the lightning, including those presented in Table 5.5. There were $n_1 = 12$ seeded storms, exhibiting an average of 19.25 cloud-to-ground lightning strikes; and $n_2 = 11$ unseeded storms, with an average of 69.45 strikes.

Inspecting the data in Table 5.5, it is apparent that the distribution of lightning counts for the unseeded storms is distinctly non-Gaussian. In particular, the set

Table 5.5

Counts of Cloud-to-Ground Lightning for Experimentally Seeded and Nonseeded Storms[a]

Seeded		Unseeded	
Date	Lightning strikes	Date	Lightning strikes
7/20/65	49	7/2/65	61
7/21/65	4	7/4/65	33
7/29/65	18	7/4/65	62
8/27/65	26	7/8/65	45
7/6/66	29	8/19/65	0
7/14/66	9	8/19/65	30
7/14/66	16	7/12/66	82
7/14/66	12	8/4/66	10
7/15/66	2	9/7/66	20
7/15/66	22	9/12/66	358
8/29/66	10	7/3/67	63
8/29/66	34		

[a] [From Baughman *et al.* (1976), reproduced with permission of the American Meteorological Society.]

contains one very strong outlier of 358 strikes. One suspects, therefore, that uncritical application of the t test [Eq. (5.8)] to test the significance of the difference in the observed mean numbers of lightning strikes could produce misleading results. This is because the single very large value of 358 strikes leads to a sample standard deviation for the unseeded storms of 98.93 strikes. This large sample standard deviation leads to the (assumed) t-distributed sampling distribution of the difference of means having a very large spread, so that even rather large values of the test statistic will be judged as being fairly ordinary.

The mechanics of applying the rank-sum test to the data in Table 5.5 are shown in Table 5.6. In the left-hand portion of the table, the 23 data points are pooled and ranked, consistent with the null hypothesis that all the data came from the same population, regardless of the labels "S" (seeded) or "N" (not seeded). In

Table 5.6

Rank-Sum Test Procedure Using the Cloud-to-Ground Lightning Data in Table 5.5[a]

Pooled data			Segregated data			
Strikes	Seeded?	Rank				
0	N	1			N	1
2	S	2	S	2		
4	S	3	S	3		
9	S	4	S	4		
10	N	5.5			N	5.5
10	S	5.5	S	5.5		
12	S	7	S	7		
16	S	8	S	8		
18	S	9	S	9		
20	N	10			N	10
22	S	11	S	11		
26	S	12	S	12		
29	S	13	S	13		
30	N	14			N	14
33	N	15			N	15
34	S	16	S	16		
45	N	17			N	17
49	S	18	S	18		
61	N	19			N	19
62	N	20			N	20
63	N	21			N	21
82	N	22			N	22
358	N	23			N	23
			Sums of ranks: $R_1 = 108.5$, $R_2 = 167.5$			

[a] In the left portion of this table, the $n_1 + n_2 = 23$ counts of lightning strikes are pooled and ranked. In the right portion of the table, the observations are segregated according to their labels of seeded (S) or not seeded (N) and the sums of the ranks for the two categories (R_1 and R_2) are computed.

the right-hand portion of the table, the data are segregated according to these labels, and the sums of the ranks of the two groups are computed. It is clear from this portion of Table 5.6 that the smaller numbers of strikes tend to be associated with the seeded storms, and the larger numbers of strikes tend to be associated with the unseeded storms. These differences are reflected in the differences in the sums of the ranks: R_1 for the seeded storms is 108.5, and R_2 for the unseeded storms is 167.5. The null hypothesis that seeding does not affect the number of lightning strikes can be rejected if this difference of this magnitude between R_1 and R_2 is sufficiently unusual against the backdrop of all possible $(23!)/[(12!)(11!)] = 1{,}352{,}078$ distinct arrangements of these data under H_0.

Consider the Mann–Whitney U statistic [Eq. (5.22)], corresponding to the sum of the ranks of the seeded data, $U_1 = 108.5 - (6)(12 + 1) = 30.5$. The null distribution of all 1,352,078 possible values of U_1 for these data is closely approximated by the Gaussian distribution having [Eq. (5.23)] $\mu_U = (12)(11)/2 = 66$ and $\sigma_U = [(12)(11)(12 + 11 + 1)/12]^{1/2} = 16.2$. Within this Gaussian distribution, the observed $U_1 = 30.5$ corresponds to a standard Gaussian $z = (30.5 - 66)/16.2 = -2.19$. Table B.1 shows the (one-tailed) p value associated with the observations to be 0.014, indicating that approximately 1.4% of the 1,352,078 possible values of U_1 under H_0 are smaller than the observed one. Accordingly, H_0 usually would be rejected. ¤

There is also a classical nonparametric test, the Wilcoxon signed-rank test, analogous to the paired two-sample parametric test of Eq. (5.11). As is the case for its parametric counterpart, the signed-rank test takes advantage of positive correlation between the pairs of data in assessing possible differences in location. In common with the unpaired rank-sum test, the signed-rank test statistic is based on ranks rather than the numerical values of the data. Therefore this test also does not depend on the distribution of the underlying data, and is resistant to outliers.

Denote the data pairs (x_i, y_i), for $i = 1, \ldots, n$. The signed-rank test is based on the set of n differences, D_i, between the n data pairs. If the null hypothesis is true, and the two data sets represent batches from the same population, roughly equally many of these differences will be positive and negative, and the overall magnitudes of the positive and negative differences should be comparable. The comparability of the positive and negative differences is assessed by ranking them in absolute value; that is, the n values of D_i are transformed to the series of ranks:

$$T_i = \text{rank}|D_i| = \text{rank}|x_i - y_i|. \tag{5.24}$$

Data pairs for which $|D_i|$ are equal are assigned the average rank of the tied values of $|D_i|$.

If the null hypothesis is true, the labeling of a given data pair as (x_i, y_i) could just as well have been reversed, so that the ith data pair was just as likely to be labeled (y_i, x_i). Changing the ordering reverses the sign of D_i, but yields the same $|D_i|$. The unique information in the pairings that were actually observed is cap-

tured by separately summing the ranks, T_i, corresponding to pairs having positive or negative values of D_i, denoting as T either the statistic

$$T^+ = \sum_{D_i>0} T_i \tag{5.25a}$$

or

$$T^- = \sum_{D_i<0} T_i, \tag{5.25b}$$

respectively. Any cases for which $D_i = 0$ are allocated equally between T^+ and T^-. Tables of null distribution probabilities sometimes require choosing the smaller of Eqs. (5.25a) and (5.25b). However, knowledge of one is sufficient for the other, since $T^+ + T^- = n(n + 1)/2$.

The null distribution of T is arrived at conceptually by considering again that H_0 implies the labeling of one or the other of each datum in a pair as x_i or y_i is arbitrary. Therefore, under the null hypothesis there are 2^n equally likely arrangements of the $2n$ data values at hand, and the resulting 2^n possible values of T constitute the relevant null distribution. As before, it is not necessary to compute all possible values of the test statistic, since for moderately large n (greater than about 20) the null distribution is approximately Gaussian, with parameters

$$\mu_T = \frac{n(n + 1)}{4} \tag{5.26a}$$

and

$$\sigma_T = \left[\frac{n(n + 1)(2n + 1)}{24}\right]^{1/2}. \tag{5.26b}$$

Under the null hypothesis, T will be close to μ_T because the numbers and magnitudes of the ranks T_i will be comparable for the negative and positive differences D_i. If there is a substantial difference between the X values and Y values in location, most of the large ranks will correspond to either the negative or positive D_i values, implying that T will be either very large or very small.

Example 5.7. Comparing Thunderstorm Frequencies Using the Signed-Rank Test

The procedure for the Wilcoxon signed-rank test is illustrated in Table 5.7. Here the paired data are counts of thunderstorms reported in the northeastern United States (X) and the Great Lakes states (Y) for the $n = 21$ years 1885–1905. Since the two areas are close geographically, one expects that large-scale flow conducive to thunderstorm formation in one of the regions would be generally conducive in the other region as well. It is thus not surprising that the reported thunderstorm counts in the two regions are substantially positively correlated.

Table 5.7

Wilcoxon Signed-Rank Test Procedure Using Data for Counts of Thunderstorms Reported in Northeastern United States (X) and Great Lakes States (Y) for 1885–1905[a]

Paired data			Differences		Segregated ranks	
Year	X	Y	D_i	Rank $\lvert D_i \rvert$	$D_i > 0$	$D_i < 0$
1885	53	70	−17	20		20
1886	54	66	−12	17.5		17.5
1887	48	82	−34	21		21
1888	46	58	−12	17.5		17.5
1889	67	78	−11	16		16
1890	75	78	− 3	4.5		4.5
1891	66	76	−10	14.5		14.5
1892	76	70	+ 6	9	9	
1893	63	73	−10	14.5		14.5
1894	67	59	+ 8	11.5	11.5	
1895	75	77	− 2	2		2
1896	62	65	− 3	4.5		4.5
1897	92	86	+ 6	9	9	
1898	78	81	− 3	4.5		4.5
1899	92	96	− 4	7		7
1900	74	73	+ 1	1	1	
1901	91	97	− 6	9		9
1902	88	75	+13	19	19	
1903	100	92	+ 8	11.5	11.5	
1904	99	96	+ 3	4.5	4.5	
1905	107	98	+ 9	13	13	
			Sums of ranks: $T^+ = 78.5$, $T^- = 152.5$			

[a] Analogously to the procedure of the rank-sum test (Table 5.6), the absolute values of the annual differences, $|D_i|$, are ranked and then segregated according to whether D_i is positive or negative. The sum of the ranks of the segregated data constitute the test statistic.

For each year the difference in reported thunderstorm counts, D_i, is computed, and the absolute values of these differences are ranked. Years having equal differences, in absolute value, are assigned the average rank (e.g., 1892, 1897, and 1901 respectively have the 8th, 9th, and 10th smallest $|D_i|$, and are all assigned the rank 9). In the final two columns, the ranks for the years with positive and negative D_i, respectively, are added; yielding $T^+ = 78.5$ and $T^- = 152.5$.

If the null hypothesis that the reported thunderstorm frequencies in the two regions are equal is true, then labeling of counts in a particular year as being northeastern or Great Lakes is arbitrary and thus so is the sign of each D_i. Consider, arbitrarily, the test statistic T as the sum of the ranks for the positive differences, $T^+ = 78.5$. Its unusualness in the context of H_0 is assessed in relation to the $2^{21} = 2{,}097{,}152$ values of T^+ that could result from all the possible permutations of the data under the null hypothesis. This null distribution is closely ap-

proximated by the Gaussian distribution having $\mu_T = (21)(22)/4 = 115.5$ and $\sigma_T = [(21)(22)(42 + 1)/24]^{1/2} = 28.77$. The p value for this test is then obtained by computing the standard Gaussian $z = (78.5 - 115.5)/28.77 = -1.29$. Assuming there was no reason to expect one or the other of the two regions to have had more reported thunderstorms, the test is two-tailed (H_A is simply "not H_0"), so the desired p value is $\Pr\{z \leq -1.29\} + \Pr\{z > +1.29\} = 2\Pr\{z \leq -1.29\} = 0.197$. The null hypothesis would not be rejected in this case. Note that the same result would be obtained if the test statistic $T^- = 152.5$ had been chosen instead. ¤

5.3.2 Resampling Tests

Since the advent of inexpensive and fast computing, another approach to nonparametric testing has become practical. This approach is based on the construction of artificial data sets from a given collection of real data, by *resampling* the observations in a manner consistent with the null hypothesis. Sometimes such methods are also known as *rerandomization tests* or *Monte-Carlo tests*. Resampling methods are highly adaptable to different testing situations, and there is considerable scope for the analyst to creatively design new tests to meet particular needs.

The basic idea is to build up a collection of artificial data batches of the same size as the actual data at hand, and then compute the test statistic of interest for each artificial batch. This procedure results in as many artificial values of the test statistic as there are artificially generated data batches. Taken together, these reference test statistics constitute an estimated null distribution against which to compare the test statistic computed from the original data.

In general, resampling tests have two very appealing advantages. The first is that no assumptions regarding an underlying theoretical distribution for the data are necessary, because the procedures consist entirely of operations on the data themselves. The second is that any statistic that may be suggested as important by the physical nature of the problem can form the basis of the test, so long as it can be computed from the data. For example, when investigating location (i.e., overall magnitudes) of a batch of data, one is not confined to the conventional tests involving the arithmetic mean, but can as easily use alternative measures such as the median, trimean, or geometric mean if any of these are more meaningful to the problem at hand. The data being tested can be scalar (each data point is one number) or vector-valued (data points are comprised of pairs, triples, etc.), as dictated by the structure of each particular problem. Resampling procedures involving vector-valued data can be especially useful when the effects of spatial correlation must be captured by the test.

A very useful technique in single-sample situations is one called the "bootstrap." It has been described with mathematical details in Efron (1982), more accessibly in Efron and Gong (1983), and popularly in Diaconis and Efron (1983).

Some examples of its use in climatology are given in Downton and Katz (1993) and Mason and Mimmack (1992). The bootstrap operates by constructing the artificial data batches using sampling *with replacement* from the original data. Conceptually, the sampling process is equivalent to writing each of the n data values on separate slips of paper and putting all n slips of paper in a hat. To construct one bootstrap sample, n slips of paper are drawn from the hat and their data values recorded, but each slip is put back in the hat and mixed (this the meaning of "with replacement") before the next slip is drawn. This process is repeated a large number, say, n_B, of times, yielding n_B bootstrap samples of size n. A particular member of the original data set may be drawn more than once, or not at all, in a given bootstrap sample.

As a practical matter, one programs a computer to do the resampling. Fundamental to this process are algorithms called *random-number generators*, or more precisely, *uniform* [0, 1] *random number generators*. Such algorithms produce streams of numbers that resemble independent values drawn randomly from the uniform [0, 1] PDF $f(u) = 1, 0 \leq u \leq 1$. This distribution is also known as the *rectangular distribution*. As the name and density function imply, these random numbers are equally likely to take on any value between 0 and 1. Press *et al.* (1986) suggest several algorithms suitable for generating uniform random numbers.

Bootstrap samples are generated by assigning numbers, from 1 to n, to each of the n observed data values to be resampled. The output of the random-number generator is then used to form the bootstrap sample by repeatedly choosing among the n values with equal probability. One way to do this is to use the ranks to number the data, denoting as $x_{(1)}$ the smallest and $x_{(n)}$ the largest of the n data values. Then, for each bootstrap sample to be constructed, n uniform random numbers u_j ($j = 1, \ldots, n$) are generated and used to draw from the pool of data values by inverting the empirical CDF of the data. That is, for each u_j a value of x is chosen according to $x = p^{-1}(u_j)$. Equivalently, one uses the rule

$$\text{Choose } x_{(i)} \text{ if } \quad \frac{i-1}{n} \leq u_j < \frac{i}{n}, \tag{5.27a}$$

which yields

$$i = \text{int}[nu_j + 1]. \tag{5.27b}$$

Here int[·] indicates retention of the integer part only, or truncation of fractions. Typically this procedure is conducted perhaps $n_B = 1000$ to $n_B = 10{,}000$ times, with the statistic of interest being computed for each of these n_B bootstrap samples.

Example 5.8. Bootstrap Confidence Intervals for a Complicated Sample Statistic

As a one-sample procedure, the bootstrap is often used to estimate confidence intervals around observed values of a test statistic. Because one does not need

Fig. 5.4 Histogram of $n_B = 10{,}000$ bootstrap estimates of the standard deviation of the logarithms of the 1933–1982 Ithaca January precipitation data. The sample standard deviation computed directly from the data is 0.537. The 95% confidence interval for the statistic, as estimated using the percentile method, is also shown.

to know the analytical form of its sampling distribution, the procedure can be applied to any test statistic, regardless of how mathematically complicated it may be. To take a hypothetical example, consider the standard deviation of the logarithms, $s_{\ln x}$, of the 1933–1982 Ithaca January precipitation data in Table A.2 of Appendix A. This statistic has been chosen for this example arbitrarily, to illustrate that any computable sample statistic can be bootstrapped. Here scalar data are used, but Efron and Gong (1983) illustrate the bootstrap using vector-valued (paired) data, for which a confidence interval of the Pearson correlation coefficient is estimated.

The value of $s_{\ln x}$ computed from the $n = 50$ data values is 0.537, but in order to make inferences about the true value, one needs to know or estimate its sampling distribution. Figure 5.4 shows a histogram of the sample standard deviations computed from $n_B = 10{,}000$ bootstrap samples of size $n = 50$ from the logarithms of this data set. The necessary computations took only a few seconds. This empirical distribution approximates the sampling distribution of the $s_{\ln x}$ for these data.

Confidence regions for $s_{\ln x}$ are most easily approached using the *percentile method* (Efron, 1982; Efron and Gong, 1983), which is a straightforward and in-

tuitive procedure. To form a $(1 - \alpha)\%$ confidence interval, one simply finds the values of the parameter estimates defining largest and smallest $n_B\ \alpha/2$ of the n_B bootstrap estimates. These values also define the central $n_B\ (1 - \alpha)$ of the estimates, which is the region of interest. In Fig. 5.4, for example, the estimated 95% confidence interval for $s_{\ln x}$ is between 0.41 and 0.65. Better and more sophisticated methods of bootstrap confidence interval construction, called *bias-corrected* (BC) intervals, are given in Efron (1982, 1987). Zwiers (1990) and Downton and Katz (1993) also sketch the mechanics of BC bootstrap confidence intervals. ¤

While the bootstrap is typically used in one-sample problems, two (or more)-sample problems can often be approached using *permutation* tests. These have been described, for example, by Mielke *et al.* (1981). Such tests operate in the same spirit as the Wilcoxon–Mann–Whitney test described previously, but with two important differences. First, any test statistic that may be meaningful may be employed, including but certainly not limited to the particular function of the ranks given in Eq. (5.22). Among other things, the lifting of restrictions on the mathematical form of possible test statistics expands the range of applicability of permutation tests to vector-valued data. For example, Mielke *et al.* (1981) provide a simple illustrative example using two batches of bivariate data ($\mathbf{x} = [x, y]$) and the Euclidian distance measure examining the tendency of the two batches to cluster in the $[x, y]$ plane. Zwiers (1987) gives an example of a permutation test that uses higher-dimensional multivariate Gaussian variates.

Again in common with the Wilcoxon–Mann–Whitney test, the null hypothesis is that the data batches are equal with respect to the test statistic of interest. Thus, under H_0 the labeling of each data point as belonging to one batch or another is arbitrary. This leads logically to the formation of the null distribution using samples drawn by computer from a pool of the combined batches of test data. If two batches of size n_1 and n_2 are to be compared, the pooled set to be resampled again contains $n = n_1 + n_2$ points. However, rather than computing the test statistic using all possible $n!/(n_1!)(n_2!)$ groupings of the pooled data, the pool is merely sampled some large number (again, perhaps 1000 or 10,000) of times. Here the samples are drawn *without replacement* (the essential distinction with respect to the bootstrap), so that each of the individual n observations is represented once and once only in one or the other of the artificial batches of size n_1 and n_2. For each of these pairs of synthetic samples the test statistic comparing the two batches is computed, and the resulting distribution of outcomes constitutes the null distribution against which the observed test statistic can be compared in the usual way.

Example 5.9. Two-Sample Permutation Test for a Complicated Sample Statistic

Consider again the lightning data in Table 5.5. Assume that their dispersion is best (from the standpoint of some criterion external to the hypothesis test) char-

acterized by the L-scale statistic (Hosking, 1990), λ_2. This statistic can be thought of as the average value of the difference of samples of size 2 drawn from the distribution underlying the data sample. This statistic can be computed from a sample of size n_1 using

$$\lambda_2 = \frac{(n_1 - 2)!}{n_1!} \sum_{i=1}^{n_1-1} \sum_{j=i+1}^{n_1} |x_i - x_j|. \tag{5.28}$$

Equation (5.28) amounts to half the average difference, in absolute value, between all possible pairs of points in the sample of size n_1. For a tightly clustered batch of data each term in the sum will be small, and therefore λ_2 will be small. For a data batch that is highly variable, some of the terms in Eq. (5.28) will be very large, and λ_2 will be correspondingly large.

To compare sample λ_2 values from the seeded and unseeded storms in Table 5.5, one would probably use either the ratio or the difference of λ_2 for the two batches. Again, the resampling test procedure provides the freedom to choose the one (or some other) making more sense for the problem at hand. Suppose that the most relevant test statistic is the ratio $[\lambda_2(\text{seeded})]/[\lambda_2(\text{unseeded})]$. According to the null hypothesis, which states that the two batches have the same L scale, this statistic should be near one. If the seeded storms are more variable with respect to the numbers of lightning strikes, the ratio statistic should be greater than one. If the seeded storms are less variable, the statistic should be less than one. The ratio of L-scales is chosen for this example arbitrarily, to illustrate that any computable function of the data can be used as the basis of a permutation test.

The null distribution of the test statistic is built up by sampling some (say, 10,000) of the $23!/(12!11!) = 1{,}352{,}078$ distinct partitions, or permutations, of the $n = 23$ data points into two batches of sizes $n_1 = 12$ and $n_2 = 11$. A modification of Eq. (5.27), to make sure that each n value is chosen once and only once, is used to draw the samples. For each partition, λ_2 is computed according to Eq. (5.28) for each of the two batches, and their ratio (with the value for the $n_1 = 12$ batch in the numerator) is computed and stored. The observed value of the ratio of the L scales, 0.188, is then evaluated with respect to this empirically generated null distribution.

Figure 5.5 shows a histogram for the null distribution of the ratios of the L-scales constructed from 10000 permutations of the original data. The observed value of 0.188 is less than all except 49 of these 10000 estimates, which would lead to the null hypothesis being soundly rejected. Notice that this null distribution has the unusual feature of being bimodal, or having two humps. This characteristic results from the single outlier in Table 5.5, 358 lightning strikes on 9/12/66, producing a very large L-scale in whichever partition it is assigned. Partitions where this observation has been assigned to the "unseeded" group are in the left hump, and those for which the outlier has been assigned to the "seeded" group are in the right hump. None of the usual parametric tests could represent this null distribution at all adequately. ¤

Fig. 5.5 Histogram for the null distribution of the ratio of the L-scales for lightning counts of seeded versus unseeded storms in Table 5.5. The observed ratio of 0.188, which is less than all except 49 of the 10000 resampling estimates of the ratio, constitutes very strong evidence that the lightning production by seeded storms was less variable than by unseeded storms. This null distribution is bimodal because the one outlier (353 strikes on 9/12/66) produces a very large L-scale in whichever of the two partitions it has been randomly assigned.

It is important to note that both bootstrap and permutation tests make sense only when the underlying data to be resampled are independent. If the data are mutually correlated (exhibiting, e.g., time correlation or persistence) the results of these testing approaches will be misleading (Zwiers, 1987, 1990). This is because the random sampling shuffles the original data, destroying the ordering that produces the correlation. Solow (1985) has suggested a way around this problem, which involves transformation of the data to an uncorrelated series using time-series methods (Chapter 8), for example, by fitting an autoregressive process. The bootstrapping or permutation test is then carried out on the transformed data, and synthetic samples exhibiting correlation properties similar to the original data can then be obtained by applying the inverse transformation.

5.4 Field Significance and Multiplicity

Special problems occur when statistical tests involving atmospheric (or other geophysical) fields must be performed. In this context the term *atmospheric field* usually connotes a two-dimensional (horizontal) array of geographic locations at which data are available. It may be, for example, that two atmospheric models (one, perhaps, reflecting an increase of the atmospheric CO_2 concentration) both produce realizations of surface temperature at each of many gridpoints, and an important question is whether the average temperatures portrayed by the two models are significantly different.

The problem of significance tests for this kind of data is often approached by conducting individual tests at each of the gridpoints using a procedure such as that in Eq. (5.8). If appropriate, a correction for serial correlation of the underlying data such as that reflected in Eq. (5.13) would be part of each of these "local" tests. Having conducted the local tests, however, it still remains to evaluate, collectively, the overall significance of the differences between the fields. This evaluation of overall significance is sometimes called determination of *global* or *pattern* significance. There are two major difficulties associated with this step. These derive from the problems of test *multiplicity*, and from spatial correlation of the underlying data.

5.4.1 The Multiplicity Problem for Independent Tests

Consider first the simpler problem of evaluating the collective significance of N independent local tests. In the context of atmospheric *field significance* testing, this situation corresponds to evaluating results from a collection of gridpoint tests when there is no spatial correlation among the underlying data at different gridpoints. Actually, the basic situation is not unique to geophysical settings, and arises any time the results of multiple, independent significance tests must be jointly evaluated. It is usually called the *problem of multiplicity*.

To fix ideas, imagine that there is a spatial array of $N = 20$ gridpoints at which local tests for the equality of means from two data sources have been conducted. These tests might be t tests, Wilcoxon–Mann–Whitney tests, or any other test appropriate to the situation at hand. From these 20 tests, it is found that $X = 3$ of them declare significant differences at the 5% level. It is sometimes naively supposed that, since 5% of 20 is 1, finding that any one of the 20 tests indicated a significant difference at the 5% level would be grounds for declaring the two fields to be significantly different, and that by extension 3 significant tests out of 20 would be very strong evidence. Although this reasoning sounds superficially plausible, it is really even approximately true only if there are very many, perhaps 1000, independent tests (see, e.g., Livezey and Chen, 1983).

Recall that declaring a significant difference at the 5% level means that, if the

null hypothesis is true and there are really no significant differences, there is a probability of 0.05 that evidence against H_0 as strong as or stronger than observed would have appeared by chance. For a single test, the situation is analogous to rolling a 20-sided die, and observing that the side with the "1" on it has come up. However, conducting $N = 20$ tests is like rolling this die 20 times: there is a substantially higher chance than 5% that the side with "1" on it comes up at least once in 20 throws, and it is this latter situation that is analogous to the simultaneous evaluation of the results from $N = 20$ independent hypothesis tests.

Thinking about this analogy between multiple tests and multiple rolls of the 20-sided die suggests that one can quantitatively analyze the multiplicity problem for independent tests in the context of the binomial distribution. In effect, one must conduct a hypothesis test on the results of the N individual independent hypothesis tests. Recall that the binomial distribution specifies probabilities for X "successes" out of N independent "trials" if the probability of "success" on any one trial is p. In the multiplicity context, X is the number of significant individual tests out of N tests conducted, and p is the level of the test.

Example 5.10. A Simple Illustration of the Multiplicity Problem

In the hypothetical example just discussed, $N = 20$ tests, $p = 0.05$ is the level of each of these tests, and $X = 3$ of the 20 tests yielded significant differences. The question of whether the means are (collectively) different at the $N = 20$ gridpoints thus reduces to evaluating $\Pr\{X \geq 3\}$, given that the null distribution for the number of significant tests is binomial with $N = 20$ and $p = 0.05$. Using the binomial PDF [Eq. (4.1)] with these two parameters, one finds $\Pr\{X = 0\} = 0.358$, $\Pr\{X = 1\} = 0.377$, and $\Pr\{X = 2\} = 0.189$. Thus, $\Pr\{X \geq 3\} = 1 - \Pr\{X < 3\} = 0.076$, and the null hypothesis that the two mean fields, as represented by the $N = 20$ gridpoints, are equal would not be rejected at the 5% level. Since $\Pr\{X = 3\} = 0.060$, finding four or more significant local tests would result in a declaration of "field significance."

Even if there are no real differences, the chances of finding at least one significant test result out of twenty are almost 2 out of 3, since $\Pr\{X = 0\} = 0.358$. Until one is aware of and accustomed to the issue of multiplicity, results such as these seem counter-intuitive. Livezey and Chen (1983) have pointed out some instances in the literature of the atmospheric sciences where a lack of awareness of the multiplicity problem has lead to conclusions that were not really supported by the data. ¤

5.4.2 Field Significance Given Spatial Correlation

Dealing with the multiplicity problem in the context of the binomial distribution, as described above, is a straightforward and satisfactory solution to the multiplicity problem when evaluating a collection of independent tests. When the tests are performed using data from spatial fields, however, the positive gridpoint-

to-gridpoint correlation of the underlying data produces statistical dependence among the local tests.

Informally, one can imagine that positive correlation between data at two locations would result in the probability of a type I error (falsely rejecting H_0) at one location being larger if a type I error had occurred at the other location. Thus, false rejections of the null hypothesis would tend to cluster in space, leading to the erroneous impression that a spatially coherent difference between the fields existed. Loosely, one can imagine that there are some number $N' < N$ of "effectively independent" gridpoint tests. Therefore, as pointed out by Livezey and Chen (1983), the binomial test of local test results to correct for the multiplicity problem provides only a lower limit on the p value pertaining to field significance. It is useful to perform that simple test, however, because a set of local tests providing insufficient evidence to reject H_0 under the assumption of independence will certainly not give a stronger result when the spatial dependence has been accounted for. In such cases there is no point in conducting the more elaborate tests required to account for the spatial dependence.

To date, most approaches to incorporating the effects of spatial dependence into tests of field significance have involved use of permutation tests or other randomization procedures. As described previously, the idea is to generate an approximation to the sampling distribution of the test statistic [in this case the number of significant local, or gridpoint tests—called the *counting norm* statistic by Zwiers (1987)] by repeated resampling of the data in a way that mimics the actual data-generation process if the null hypothesis is valid. Usually the null hypothesis specifies that there are no significant differences in the field of tests. Different testing situations present different challenges, and considerable creativity is sometimes required for the analyst to devise a consistent resampling procedure. Further discussion on this topic can be found in Livezey and Chen (1983), Livezey (1985), Preisendorfer and Barnett (1983), Wigley and Santer (1990), and Zwiers (1987).

Example 5.11. The Multiplicity Problem with Spatial Correlation

An instructive example of the use of a permutation test to assess the joint results of a field of hypothesis tests is presented by Livezey and Chen (1983), using data from Chen (1982b). The basic problem is illustrated in Fig. 5.6a, which shows the field of correlations between Northern Hemisphere winter (December–February) 700-mb heights, and values of the Southern Oscillation Index (SOI) (Fig. 3.10) for the previous summer (June–August). The areas of large positive and negative correlation suggest that the SOI might be useful as one element of a long-range (6 months ahead) forecast procedure for winter weather. First, however, a formal test that the field of correlations in Fig. 5.6a is different from zero is in order.

The testing process begins with individual significance tests for the correlation coefficients at each gridpoint. If the underlying data (here the SOI values and 700-mb heights) approximately follow Gaussian distributions with negligible year-to-year correlation, an easy approach to this suite of tests is to use the Fisher z

Fig. 5.6 Correlations of Northern Hemisphere winter (December–February) 700-mb heights with (a) the Southern Oscillation Index for the previous summer (June–August) and (b) one particular set of independent Gaussian random numbers. Shaded areas show positive correlation, and the contour interval is 0.1. The strong spatial correlation of the 700-mb-height field produces the spatially coherent correlation with the random-number series as an artifact, complicating interpretation of gridpoint significance tests. [From Chen (1982b), reproduced with permission of the American Meteorological Society.]

transformation,

$$\mathfrak{Z} = \frac{1}{2} \ln\left[\frac{1 + r}{1 - r}\right], \tag{5.29}$$

where r is the Pearson correlation [Eq. (3.17)]. Under the null hypothesis that the correlation r is zero,the distribution of $\mathfrak{Z}$ approximates the Gaussian distribution with $\mu = 0$ and $\sigma = (n - 3)^{-1/2}$. (If a different null hypothesis were appropriate,

b

CORRELATION BETWEEN NOISE AND DJF 700 MB HEIGHT

Fig. 5.6 *Continued.*

the mean of the corresponding Gaussian distribution would be the $\mathfrak{Z}$ transform of the correlation under that null hypothesis.) Since Chen (1982b) used $n = 29$ years, values of $\mathfrak{Z}$ larger in absolute value than $1.96/\sqrt{26} = 0.384$, corresponding to correlations larger than 0.366 in absolute value, would be declared (locally) significant at the 5% level. A sufficiently large area in Fig. 5.6a exhibits correlations of these magnitudes that the preliminary (binomial) test, accounting for the multiplicity problem, rejects the null hypothesis of zero correlation under the assumption that the local tests are mutually independent.

The need for a more sophisticated test, accounting also for the effects of spatial correlation of the 700-mb heights, is underscored by the correlation field in Fig. 5.6b. This shows the correlations between the same 29-year record of Northern Hemisphere 700-mb heights with 29 independent Gaussian random numbers, i.e., a random series similar to Fig. 5.2a. Clearly the real correlation of 700-mb

heights with this series of random numbers is zero, but the substantial spatial correlation among the 700-mb heights yields spatially coherent areas of chance sample correlations that are deceptively high.

The approach to this particular problem taken by Livezey and Chen (1983) was to repeatedly generate sequences of 29 independent Gaussian random variables, as a null hypothesis stand-in for the observed series of SOI values, and tabulate the numbers of local tests (erroneously) rejecting H_0 for each sequence. This is an appropriate design, since under H_0 there is no real correlation between the 700-mb heights and the SOI, and it is the spatial correlation of the 700-mb heights that must be preserved in order to simulate the true data-generating process. Another approaches could have been to repeatedly use the 29 observed SOI values, but in random orders, in place of the Gaussian random numbers. Alternatively, sequences of correlated values generated from a time-series model (Chapter 8) mimicking the SOI could have been used.

Livezey and Chen chose to resample the distribution of numbers of rejected tests 200 times, with the resulting 200 counts of significant local tests constituting

Fig. 5.7 Null distribution for the percentage of hemispheric area exhibiting significant local tests for correlation between Northern Hemisphere winter 700-mb heights and series of Gaussian random numbers. Numbers of significant local tests for correlations between the heights and concurrent SOI values, the SOI for the previous fall, and the SOI for the previous summer are indicated as "LAG 0," "LAG 1," and "LAG 2," respectively. The result for correlations with summer SOI are inconsistent with the null hypothesis of no relationship between the height field and the SOI, which would be rejected. Results for LAG 0 (i.e., contemporaneous) correlation are not strong enough to reject this H_0. [From Livezey and Chen (1983), reproduced with permission of the American Meteorological Society.]

the relevant null distribution. A histogram of this distribution is shown in Fig. 5.7, with the largest 5% of the values shaded. The number of local tests from Fig. 5.6a that were found to be significant are indicated by the arrow labeled "LAG 2" (2 seasons ahead of winter). Clearly the results for correlation with the summer SOI are located well out on the tail, and are therefore judged to be highly significant. Results for correlation with fall (September–November) are indicated by the arrow labeled "LAG 1," and are evidently significantly different from zero at the 5% level, but less strongly so than for the summer SOI observations. The corresponding results for contemporaneous (winter SOI and winter 700-mb heights) correlations, by contrast, do not fall in the most extreme 5% of the null distribution values, and H_0 would not be rejected for these data. ¤

Exercises

5.1. For the June temperature data in Table A.3:
   (a) Use a two-sample t test to investigate whether the average June temperatures in El Niño and non–El Niño years are significantly different. Assume the Gaussian distribution is an adequate approximation to the distribution of the test statistic.
   (b) Construct a 95% confidence interval for the difference in average June temperature between El Niño and non–El Niño years.

5.2. Calculate n', the equivalent number of independent samples, for the two sets of minimum air temperature in Table A.1.

5.3. Use the data set in Table A.1 in Appendix A to test the null hypothesis that the average minimum temperatures for Ithaca and Canandaigua in January 1987 are equal. Compute p values, assuming the Gaussian distribution is an adequate approximation to the null distribution of the test statistic, and
   (a) H_A = the minimum temperatures are different for the two locations.
   (b) H_A = the Canandaigua minimum temperatures are warmer.

5.4. Given that the correlations in Fig. 5.6a were computed using 29 years of data, use the Fisher $\mathfrak{Z}$ transformation to compute the magnitude of the correlation coefficient that was necessary for the null hypothesis to be rejected at the 5% level, versus the alternative that $r \neq 0$.

5.5. Test the fit of the Gaussian distribution to the July precipitation data in Table 4.5, using
   (a) A K-S test
   (b) A Chi-square (χ^2) test

5.6. Test whether the 1951–1980 July precipitation data in Table 4.5 could have been drawn from the same distribution as the 1951–1980 January precipitation comprising part of Table A.2, using the likelihood ratio test, assuming gamma distributions.

5.7. Use the Wilcoxon–Mann–Whitney test to investigate whether the magnitudes of the pressure data in Table A.3 are lower in El Niño years:

(a) Using the exact one-tailed critical values 18, 14, 11, and 8 for tests at the 5%, 2.5%, 1%, and 0.5% levels, respectively, for the smaller of U_1 and U_2.

(b) Using the Gaussian approximation to the sampling distribution of U (even though the sample size is too small for the approximation to be good).

5.8. Discuss how the sampling distribution of the skewness coefficient [Eq. (3.8)] of June precipitation at Guayaquil could be estimated using the data in Table A.3, by bootstrapping. How could the resulting bootstrap distribution be used to estimate a 95% confidence interval for this statistic? If the appropriate computing resources are available, implement your algorithm.

5.9. Discuss how to construct a resampling test to investigate whether the variance of June precipitation at Guayaquil is different in El Niño versus non–El Niño years, using the data in Table A.3. If the appropriate computing resources are available, implement your algorithm.

5.10. A published article contains a statistical analysis of historical summer precipitation data in relation to summer temperatures using individual t tests for 121 locations at the 10% level. The study investigates the null hypothesis of no difference in total precipitation between the 10 warmest summers in the period 1900–1969 and the remaining 60 summers, reports that 15 of the 121 tests exhibit significant results, and claims that the overall pattern is therefore significant. Evaluate this claim.

Chapter 6 Statistical Weather Forecasting

6.1 Background

Much of operational weather and long-range ("climate") forecasting has a statistical basis. As a nonlinear dynamical system, the atmosphere is not perfectly predictable in a deterministic sense. Consequently, statistical methods are useful, and indeed necessary, parts of the forecasting enterprise.

Some statistical forecast methods operate without information from the fluid-dynamical *numerical weather prediction* (NWP) models that have become the mainstay of weather forecasting for lead times ranging from one to a few days in advance. Such "pure" statistical forecast methods are sometimes referred to as *classical*, reflecting their prominence in the years before NWP information was available. These methods are still viable and useful at very short lead times (hours in advance), or very long lead times (weeks or more in advance), for which NWP information is either not available with sufficient promptness or accuracy, respectively.

Another important application of statistical methods to weather forecasting is in conjunction with NWP information. Statistical forecast models are routinely used to enhance the results of dynamical forecasts at operational weather forecasting centers throughout the world, and are essential as *guidance* products to aid weather forecasters. The combined statistical and dynamical approaches are especially important for providing forecasts for quantities (e.g., probabilities) or locations (e.g., particular cities rather than gridpoints) not represented by the NWP models.

The types of statistical forecasts mentioned so far are *objective*, in the sense that a given set of inputs always produces the same particular output. However, another important aspect of statistical weather forecasting is in the subjective formulation of forecasts, particularly when the forecast quantity is a probability or set of probabilities. Here the Bayesian interpretation of probability as a quantified degree of belief is fundamental. Subjective probability assessment forms the basis of many operationally important forecasts, and is a technique that could be used more broadly to enhance the information content of operational forecasts.

6.2 Review of Least-Squares Regression

Much of statistical weather forecasting is based on the statistical procedure known as *least-squares regression.* In this section, the fundamentals of least-squares regression are reviewed. Much more complete treatments can be found in Draper and Smith (1981) and Neter *et al.* (1985).

6.2.1 Simple Linear Regression

Regression is most easily understood in the case of "simple" linear regression, which describes the linear relationship between two variables, say, x and y. Conventionally the symbol x is used for the "independent," or *predictor* variable, and the symbol y is used for the "dependent" variable, or the *predictand.* Very often, more than one predictor variable is required in practical forecast problems, but the ideas for simple linear regression generalize easily to this more complex case of multiple linear regression. Therefore, most of the important ideas about regression can be presented in this section.

Essentially, simple linear regression seeks to summarize the relationship between two variables, shown graphically in their scatterplot, by a single straight line. The regression procedure chooses that line producing the *least error* for predictions of y given observations of x. Exactly what constitutes least error can be open to interpretation, but the most usual error criterion is minimization of the sum (or, equivalently, the average) of the squared errors. It is the choice of the squared-error criterion that is the basis of the term *least-squares regression.*

Choosing the squared-error criterion is conventional not because it is necessarily best, but rather because it makes the mathematics analytically tractable. Adopting the squared-error criterion results in the line-fitting procedure being fairly tolerant of small discrepancies between the line and the points. However, the fitted line will adjust substantially to avoid very large discrepancies. It is thus not resistant to outliers.

Figure 6.1 illustrates the situation. Given a data set of (x, y) pairs, the problem is to find the particular straight line

$$\hat{y} = a + bx, \tag{6.1}$$

minimizing the squared vertical distances (dotted lines) between it and the data points. The circumflex ("hat") accent indicates that the equation specifies a predicted value of y. The inset in Fig. 6.1 indicates that the vertical distances between the data points and the line, also called *errors* or *residuals*, are defined as

$$e_i = y_i - \hat{y}(x_i). \tag{6.2}$$

There is a separate residual e_i for each data pair (x_i, y_i). Note that the sign convention implied by Eq. (6.2) is for points above the line to be regarded as "posi-

Fig. 6.1 Schematic illustration of simple linear regression. The regression line, $\hat{y} = a + bx$, is chosen as the one minimizing some measure of the vertical differences (the residuals) between the points and the line. In least-squares regression that measure is the sum of the squared vertical distances. Inset shows the residual, e, as the difference between the data point and the regression line.

tive" errors, and points below the line to be "negative" errors. This is the usual convention in statistics, but is opposite to what is often seen in the atmospheric sciences, where forecasts lower than the observations (the line being below the point) are regarded as having negative errors, and vice versa. However, the sign convention for the residuals is unimportant, since it is the minimization of the sum of squared residuals that defines the best-fitting line. Combining Eqs. (6.1) and (6.2) yields the *regression equation*

$$y_i = \hat{y}_i + e_i = a + bx_i + e_i, \tag{6.3}$$

which says that the true value of the predictand is the sum of the predicted value [Equation (6.1)] and the residual.

Finding analytic expressions for the least-squares intercept, a, and the slope, b, is a straightforward exercise in calculus. In order to minimize the sum of squared residuals

$$\sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - [a + bx_i])^2, \tag{6.4}$$

it is only necessary to set the derivatives of Eq. (6.4) with respect to the parameters a and b to zero and solve. These derivatives are

$$\frac{\partial \sum_{i=1}^{n} (e_i)^2}{\partial a} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial a}$$

$$= -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \tag{6.5a}$$

and

$$\frac{\partial \sum_{i=1}^{n} (e_i)^2}{\partial b} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial b} \tag{6.5b}$$

$$= -2 \sum_{i=1}^{n} [x_i(y_i - a - bx_i)] = 0.$$

Rearrangement of Eqs. (6.5) leads to the so-called *normal equations*

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \tag{6.6a}$$

and

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} (x_i)^2. \tag{6.6b}$$

Dividing Eq. (6.6a) by n implies that the fitted regression line must pass through the point located by the two sample means of x and y. Finally, solving the normal equations for the regression parameters yields

$$b = \frac{\Sigma_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} = \frac{n \Sigma_{i=1}^{n} x_i y_i - \Sigma_{i=1}^{n} x_i \Sigma_{i=1}^{n} y_i}{n \Sigma_{i=1}^{n} (x_i)^2 - (\Sigma_{i=1}^{n} x_i)^2} \tag{6.7a}$$

and

$$a - \bar{y} - b\bar{x}. \tag{6.7b}$$

Equation (6.7a), for the slope, is similar in form to the Pearson correlation coefficient, and can be obtained with a single pass through the data using the computational form given as the second equality. Note that, as was the case for the computational form for the correlation coefficient, careless use of the computational form of Eq. (6.7a) can lead to roundoff errors since the numerator is generally the difference between two large numbers.

6.2.2 Distribution of the Residuals

Thus far, fitting the straight line has involved no statistical ideas at all. All that has been required was to define least error to mean minimum squared error. The rest has followed from straightforward mathematical manipulation of the data, namely, the (x, y) pairs. To bring in statistical ideas, it is conventional to assume that the quantities e_i are independent random variables with zero mean and constant variance. Often, the additional assumption is made that these residuals follow a Gaussian distribution.

Assuming that the residuals have zero mean is not at all problematic. In fact, one convenient property of the least-squares fitting procedure is the guarantee that

$$\sum_{i=1}^{n} e_i = 0, \tag{6.8}$$

from which it is clear that the sample mean of the residuals (dividing this equation by n) is also zero.

Imagining that the residuals can be characterized in terms of a variance is really the point at which statistical ideas begin to come into the regression framework. Implicit in their possessing a variance is the idea that the residuals scatter randomly about some mean value [recall Eq. (4.15) or (3.5)]. Equation (6.8) guarantees that the mean value around which they will scatter is zero, so it is the regression line around which the data points will scatter. One then needs to imagine a series of distributions of the residuals *conditional* on the x values, with each observed residual regarded as having been drawn from one of these conditional distributions. The constant-variance assumption really means that the variance of the residuals is constant in x, or that all the conditional distributions of the residuals have the same variance. Therefore a given residual (positive or negative, large or small) is by assumption equally likely to occur at any part of the regression line.

Figure 6.2 is a schematic illustration of the idea of a suite of conditional distributions around the regression line. The three distributions are identical, except that their means are shifted higher or lower depending on the level of the regression line (predicted value of y) for each x. Also shown in Fig. 6.2 is a schematic representation of the unconditional distribution of the predictand, y. The distributions of residuals are less spread out (have smaller variance) than the distribution of y, reflecting less uncertainty about y if a corresponding x value is known.

Central to the making of statistical inferences in the regression setting is the estimation of this (constant) residual variance from the sample of residuals. Since the sample average of the residuals is guaranteed by Eq. (6.8) to be zero, the square of Eq. (3.5) becomes

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2, \tag{6.9}$$

Fig. 6.2 Schematic illustration of distributions of residuals around the regression line, conditional on values of the predictor variable, x. The actual residuals are regarded as having been drawn from these distributions.

where the sum of squared residuals is divided by $n - 2$ because two parameters (a and b) have been estimated. Substituting Eq. (6.2) then yields

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} [y_i - \hat{y}(x_i)]^2. \tag{6.10}$$

Rather than compute the estimated residual variance using Eq.(6.10), however, it is more usual to use a computational form based on the relationship, proved in most regression texts:

$$\text{SST} = \text{SSR} + \text{SSE}. \tag{6.11}$$

The notation in Eq. (6.11) consists of acronyms describing respectively the variation in the predictand, y, and a partitioning of that variation between the portion represented by the regression and the unrepresented portion ascribed to the variation of the residuals. The term SST is an acronym for *total sum of squares* (or sum of squares, total), which has the mathematical meaning of the sum of squared deviations of the Y values around their mean,

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2. \tag{6.12}$$

This term is proportional to (by a factor of $n - 1$) the sample variance of y, and thus measures the overall variability of the predictand. The term SSR stands for the *regression sum of squares*, or the sum of squared differences between the regression predictions and the sample mean of y

$$\mathrm{SSR} = \sum_{i=1}^{n} [\hat{y}(x_i) - \bar{y}]^2, \tag{6.13a}$$

which relates to the regression equation according to

$$\mathrm{SSR} = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = b^2\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]. \tag{6.13b}$$

Equations (6.13) indicate that a regression line differing little from the sample mean of the y values will have a small slope and produce a very small SSR, while one with a large slope will exhibit some large differences from the sample mean of the predictand and therefore produce a large SSR. Finally, SSE refers to the sum of squared differences between the residuals and their mean, which is zero, or *sum of squared errors*:

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2. \tag{6.14}$$

Since this differs from Eq. (6.9) only by a factor of $n - 2$, rearrangement of Eq. (6.11) yields the computational form

$$\begin{aligned} s_e^2 &= \frac{1}{n-2}\{\mathrm{SST} - \mathrm{SSR}\} \\ &= \frac{1}{n-2}\left\{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b^2\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]\right\}. \end{aligned} \tag{6.15}$$

6.2.3 The Analysis-of-Variance Table

In practice, regression analysis is now almost universally done using "canned" computer packages. A central part of the regression output of such packages is a summary of the foregoing information in an *analysis of variance* (ANOVA) table. Usually, not all of the information in an ANOVA table will be of interest, but it is such a universal form of regression output that one should understand its components. Table 6.1 outlines the arrangement of an ANOVA table for simple linear regression, and indicates where the quantities described in the previous section are reported. The three rows correspond to the partition of the variation of the predictand as expressed in Eq. (6.11). Accordingly, the "Regression" and "Residual" entries in the "df" (degrees of freedom) and "SS" (sum of squares) columns will sum to the corresponding entry in the "Total" row. Therefore, the

Table 6.1

Generic Analysis of Variance (ANOVA) Table for Simple Linear Regression[a]

Source	df	SS	MS
Total	$n-1$	SST (6.12)	
Regression	1	SSR (6.13)	MSR = SSR / 1 (F = MSR/MSE)
Residual	$n-2$	SSE (6.14)	MSE = s_e^2

[a] The column headings df, SS, and MS stand for degrees of freedom, sum of squares, and mean square, respectively. Regression df = 1 is particular to simple (i.e., a single predictor x) linear regression. Parenthetical references are to equation numbers in the text.

ANOVA table contains some redundant information, and as a consequence the output from some regression packages will omit the "Total" row entirely.

The entries in the "MS" (mean squared) column are given by the corresponding quotients of SS/df. For simple linear regression, the regression df = 1, and SSR = MSR. Comparing this with Eq. 6.15, it can be seen that the mean-squared error (MSE) is the sample variance of the residuals. The total mean square, left blank in Table 6.1 and in the output of most regression packages, would be SST/ (n − 1), or simply the sample variance of the predictand.

6.2.4 Goodness-of-Fit Measures

The ANOVA table also presents (or allows computation of) three related measures of the "fit" of a regression, or the correspondence between the regression line and a scatterplot of the data. The first of these is the MSE. From the standpoint of forecasting, the MSE is perhaps the most fundamental of the three measures, since it indicates the variability of the observed y values (the quantities being forecast) around the forecast regression line. As such, it directly reflects the accuracy of the resulting forecasts. Referring again to Fig. 6.2, since MSE = s_e^2 this quantity indicates the degree to which the distributions of residuals cluster tightly (small MSE), or spread widely (large MSE) around a regression line. In the limit of a perfect linear relationship between x and y, the regression line coincides exactly with all the point pairs, the residuals are all zero, SST will equal SSR, SSE will be zero, and the variance of the residual distributions is also zero. In the opposite limit of absolutely no linear relationship between x and y, the regression slope will be zero, the SSR will be zero, SSE will equal SST, and the MSE will very nearly equal the sample variance of the predictand itself. In this unfortunate case, the three conditional distributions in Fig. 6.2 would be indistinguishable from the unconditional distribution of y.

The relationship of the MSE to the strength of the regression fit is also illustrated in Fig. 6.3. Panel a shows the case of a reasonably good regression, with the scatter of points around the regression line being fairly small. Here SSR and SST

Fig. 6.3 Illustration of the distinction between a fairly good regression relationship (a) and an essentially useless relationship (b). The points in panel a cluster closely around the regression line (solid), indicating small MSE, and the line deviates strongly from the average value of the predictand (dotted), producing a large SSR. In panel b the scatter around the regression line is large, and the regression line is almost indistinguishable from the mean of the predictand.

are nearly the same. Panel b shows an essentially useless regression for values of the predictand spanning the same range as in panel a. In this case the SSR is nearly zero since the regression has nearly zero slope, and the MSE is essentially the same as the sample variance of the y values themselves.

The second usual measure of the fit of a regression is the *coefficient of determination*, or R^2. This can be computed from

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{6.16}$$

and it is often also displayed as part of standard regression output. Qualitatively, the R^2 can be interpreted as the proportion of the variation of the predictand (proportional to SST) that is "described" or "accounted for" by the regression (SSR). Sometimes one sees this concept expressed as the proportion of variation "explained," although this claim is misleading: A regression analysis can quantify the nature and strength of a relationship between two variables, but can say nothing about which variable (if either) causes the other. This is the same caveat offered in the discussion of the correlation coefficient in Chapter 3. For the case of simple linear regression, the square root of the coefficient of determination is exactly (the absolute value of) the Pearson correlation between x and y.

For a perfect regression, SSR = SST and SSE = 0, so $R^2 = 1$. For a completely useless regression, SSR = 0 and SSE = SST, so that $R^2 = 0$. Again, Fig. 6.3b shows something close to this latter case. Comparison with Eq. (6.13a) reveals that the least-squares regression line is almost indistinguishable from the sample

mean of the predictand, so SSR is very small. In other words, little of the variation in y can be ascribed to the regression, so the proportion SSR/SST is nearly zero.

The third commonly used measure of the strength of the regression is the F ratio, generally given in the last column of the ANOVA table. The ratio MSR/MSE increases with the strength of the regression, since a strong relationship between x and y will produce a large MSR and a small MSE. Assuming that the residuals are independent and follow the same Gaussian distribution, and under the null hypothesis no real linear relationship, the sampling distribution of the F ratio has a known parametric form. While this distribution forms the basis of a test that is applicable in the case of simple linear regression, in the more general case of multiple regression (more than one x variable) problems of multiplicity, discussed later, invalidate it. However, even if the F ratio cannot be used for quantitative statistical inference, it is still a valid qualitative index of the strength of a regression. See, for example, Draper and Smith (1981) or Neter *et al.* (1985), for discussions of the F test for overall significance of the regression.

6.2.5 Sampling Distributions of the Regression Coefficients

Another important use of the estimated residual variance is to obtain estimates of the sampling distributions of the regression coefficients. As statistics computed from a finite set of data subject to sampling variations, the computed regression intercept and slope, a and b, also exhibit sampling variability. Estimation of their sampling distributions allows construction of confidence intervals around the particular parameter values obtained, and provides a basis for hypothesis tests about the corresponding population values.

Under the assumptions concerning the residuals listed previously, the sampling distributions for both intercept and slope are Gaussian. For the intercept the sampling distribution has parameters

$$\mu_a = a \tag{6.17a}$$

and

$$\sigma_a = s_e \left[\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{1/2}. \tag{6.17b}$$

For the slope the parameters of the sampling distribution are

$$\mu_b = b \tag{6.18a}$$

and

$$\sigma_b = \frac{s_e}{[\sum_{i=1}^{n} (x_i - \bar{x})^2]^{1/2}}. \tag{6.18b}$$

Equations (6.17a) and (6.18a) indicate that the least-squares regression parameter estimates are unbiased. Equations (6.17b) and (6.18b) show that the precision with

which the intercept and slope can be estimated from the data depend directly on the estimated standard deviation of the residuals, s_e, which is the square root of the MSE from the ANOVA table (Table 6.1). Equation (6.18b) is valid only for simple linear regression. With more than one predictor variable, an analogous (vector) equation must be used.

The output from regression packages will almost always include the standard errors [Eqs. (6.17b) and (6.18b)] in addition to the parameter estimates themselves. Some packages also include the ratios of the estimated parameters to their standard errors in a column labeled "t ratio." When this is done, a one-sample t test [Eq. (5.3)] is implied, with the null hypothesis being that the underlying ("population") mean for the parameter is zero. Sometimes a p-value associated with this test is also automatically included in the regression output.

For the case of the regression slope, this implicit t test bears directly on the meaningfulness of the fitted regression. If the estimated slope is small enough that its true value could plausibly (with respect to its sampling distribution) be zero, then the regression is not informative, or useful for forecasting. If the slope is actually zero, then the value of the predictand specified by the regression equation is always the same, and equal to its sample mean. If the assumptions are satisfied, one would reject this null hypothesis at the 5% level if the estimated slope is, roughly, at least twice as large (in absolute value) as its standard error.

The same hypothesis test for the regression intercept is often offered by computerized statistical packages as well. Depending on the problem at hand, however, this test for the intercept may or may not be meaningful. Again, the "t ratio" is just the parameter estimate divided by its standard error, so the implicit null hypothesis is that the true intercept is zero. Occasionally, this null hypothesis is physically meaningful, and if so the test statistic for the intercept is worth looking at. On the other hand, it often happens that there is no physical reason to expect that the intercept might be zero. It may be that a zero intercept is even physically impossible. In these cases this portion of the automatically generated computer output is meaningless.

Example 6.1. A Simple Linear Regression

To concretely illustrate simple linear regression, consider the January 1987 minimum temperatures at Ithaca and Canandaigua from Table A.1 in Appendix A. Let the predictor variable, x, be the Ithaca minimum temperature, and the predictand, y, be the Canandaigua minimum temperature. The scatterplot of these data is the middle panel in the bottom row of the scatterplot matrix in Fig. 3.18. A strong, positive, and reasonably linear relationship is indicated.

Table 6.2 shows what the output from a typical statistical computer package would look like for this regression. The data set is small enough that the computational formulas presented above can be worked through to verify the results. (A little work with a hand calculator will verify that $\Sigma x = 403$; $\Sigma y = 627$, $\Sigma x^2 = 10{,}803$; $\Sigma y^2 = 15{,}009$; and $\Sigma xy = 11{,}475$.) The upper portion of Table 6.2 cor-

Table 6.2

Example Output Typical of That Produced by Computer Statistical Packages, for Prediction of Canandaigua Minimum Temperature (y) Using Ithaca Minimum Temperature (x) from the January 1987 Data Set in Table A.1

Source	df	SS	MS	F
Total	30	2327.419		
Regression	1	1985.798	1985.798	168.57
Residual	29	341.622	11.780	
Variable	*Coefficient*	*s.e.*	*t-ratio*	
Constant	12.4595	0.8590	14.504	
IthacaMin	0.5974	0.0460	12.987	

responds to the template in Table 6.1, with the relevant numbers filled in. Of particular importance is MSE = 11.780, yielding as its square root the estimated sample standard deviation for the residuals, $s_e = 3.43°$ F. This standard deviation addresses directly the precision of specifying the Canandaigua temperatures on the basis of the Ithaca temperatures, since one expects about 95% of the actual predictand values to be within $\pm 2s_e = \pm 6.9°$ F of the temperatures given by the regression (cf. Table B.1). The coefficient of determination is easily computed as $R^2 = 1985.798/2327.419 = 85.3\%$. The Pearson correlation is $\sqrt{0.853} = 0.924$, as was given in Table 3.2. The value of the F statistic is very high, considering that the 99th percentile of its distribution under the null hypothesis of no real relationship is about 7.5. One could also compute the sample variance of the predictand, which would be the "total mean-square" cell of the table, as 2327.419/30 = 77.58° F^2.

The lower portion of Table 6.2 gives the regression parameters, a and b, their standard errors, and the ratios of these parameter estimates to their standard errors. The specific regression equation for this data set, corresponding to Eq. (6.1), would be

$$T_{\text{Can.}} = \underset{(0.859)}{12.46} + \underset{(0.046)}{0.597} T_{\text{Ith.}}. \tag{6.19}$$

Thus, the Canandaigua temperature would be estimated by multiplying the Ithaca temperature by 0.597 and adding 12.46° F. The intercept $a = 12.46°$ F has no special physical significance except as the predicted Canandaigua temperature when the Ithaca temperature is 0° F. Notice that the standard errors of the two coefficients have been written parenthetically below the coefficients themselves. While this is not a universal practice, it is very informative to someone reading Eq. (6.19) without the benefit of the information in Table 6.2. In particular, it allows the reader to get a qualitative sense for the significance of the slope (i.e., the parameter b). Since the estimated slope is about 13 times larger than its

standard error, it is almost certainly not really zero. This conclusion speaks directly to the question of meaningfulness of the fitted regression. On the other hand, the corresponding implied significance test for the intercept would be much less interesting, unless the possibility of a zero intercept would in itself be meaningful. ¤

6.2.6 Examining Residuals

It is not sufficient to blindly feed data to a computer regression package and uncritically accept the results. Some of the results can be misleading if the assumptions underlying the computations are not satisfied. Since these assumptions pertain to the residuals, it is important to examine the residuals for consistency with the assumptions made about their behavior.

One easy and fundamental check on the residuals can be made by examining a scatterplot of the residuals as a function of the predicted value $\hat{y}$. Many statistical computer packages provide this capability as a standard regression option. Figure 6.4a shows a scatterplot of a hypothetical data set, with the least-squares regression line, and Fig. (6.4b) shows a plot of the resulting residuals as a function of the predicted values. The residual plot presents the impression of "fanning," or exhibition of increasing spread as $\hat{y}$ increases. That is, the variance of the residuals appears to increase as the predicted value increases. This condition of nonconstant residual variance is called *heteroscedasticity*. Since the computer program that fit the regression has assumed constant residual variance, the MSE given in the ANOVA table is an overestimate for smaller values of x and y (where the points cluster close to the regression line), and an underestimate of the residual variance for larger values of x and y (where the points tend to be far from the regression line). If the regression is used as a forecasting tool, one would be overconfident about forecasts for larger values of y, and underconfident about forecasts for

Fig. 6.4 Hypothetical linear regression (a) and plot of the resulting residuals against the predicted values (b) for a case where the variance of the residuals is not constant. The scatter around the regression line in panel a increases for larger values of x and y, producing a visual impression of "fanning" in the residual plot (b). A transformation of the predictand is indicated.

Fig. 6.5 Scatterplot with regression (a) and resulting residual plot (b) for the same data in Fig. 6.4 after logarithmically transforming the predictand. The visual impression of a horizontal band in the residual plot supports the assumption of constant variance of the residuals.

smaller values of y. In addition, the sampling distributions of the regression parameters will be more variable than implied by Eqs. (6.17) and (6.18). That is, the parameters will not have been estimated as precisely as the standard regression output would lead one to believe.

Often nonconstancy of residual variance of the sort shown in Fig. 6.4b can be remedied by transforming the predictand, perhaps by using a power transformation [Eqs. (3.14)]. Figure 6.5 shows the regression and residual plots for the same data as in Fig. 6.4 after logarithmically transforming the predictand. Recall that the logarithmic transformation reduces all the data values, but reduces the larger values more strongly than the smaller ones. Thus, the long right tail of the predictand has been "pulled in" relative to the shorter left tail, as in Fig. 3.8. As a result, the transformed data points appear to cluster more evenly around the new regression line. Instead of fanning, the residual plot in Fig. 6.5b gives the visual impression of a horizontal band, indicating appropriately constant variance of the residuals (*homoscedasticity*). Note that if the fanning in Fig. 6.4b had been in the opposite sense, with greater residual variability for smaller values of $\hat{y}$ and lesser residual variability for larger values of $\hat{y}$, a transformation that "stretches" the right tail relative to the left tail (e.g., y^2) would have been appropriate.

A graphical impression of whether the residuals follow a Gaussian distribution can be obtained through a quantile–quantile (Q–Q) plot. The capacity to make these plots is also often a standard option in statistical computer packages. Figures 6.6a and 6.6b show Q–Q plots for the residuals in Figs. 6.4b and 6.5b, respectively. The residuals are plotted on the vertical, and the standard Gaussian variables corresponding to the empirical cumulative probability of each residual are plotted on the horizontal. The curvature apparent in Fig. 6.6a indicates that the residuals from the regression involving the untransformed y are positively skewed relative to the (symmetrical) Gaussian distribution. The Q–Q plot of residuals from the regression involving the logarithmically transformed y is very nearly

Fig. 6.6 Gaussian quantile–quantile plots of the residuals for the untransformed y in Fig. 6.4a (a) and the logarithmically transformed y in Fig. 6.5b (b). In addition to producing essentially constant residual variance, the logarithmic transformation of the predictand has rendered the distribution of the residuals essentially Gaussian.

linear. Evidently the logarithmic transformation has produced residuals that are close to Gaussian, in addition to stabilizing the residual variances.

Note that, if the relationship shown in Fig. 6.5a were to be used in forecasting, dimensional values of the predictand would need to be recovered in order to make the forecasts meaningful. That is, the predictand $\ln(y)$ would need to be back-transformed, yielding the forecast $y = \exp[\ln(y)] = \exp[a + bx]$.

It is also possible to investigate the degree to which the residuals are mutually independent. This is of particular interest when the underlying data are serially correlated, which is a common condition for atmospheric variables. A simple graphical evaluation can be obtained by plotting the regression residuals as a function of time. If groups of positive and negative residuals tend to cluster together (qualitatively resembling Fig. 5.2b) rather than occurring more irregularly (as in Fig. 5.2a), then time correlation can be suspected.

A popular formal test for serial correlation of regression residuals, included in many computer regression packages, is the Durbin–Watson test. This test examines the null hypothesis that the residuals are serially independent against the alternative that they are consistent with a first-order autoregressive process (see Chapter 8). The Durbin–Watson test statistic,

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \tag{6.20}$$

computes the squared differences between pairs of consecutive residuals, divided by a scaling factor. If the residuals are positively correlated, adjacent residuals will tend to be similar in magnitude, so the Durbin–Watson statistic will be relatively small. If the residuals are randomly distributed in time, the sum in the nu-

Fig. 6.7 Graphs of the 5% critical values for the Durbin–Watson statistic as a function of the sample size, for $K = 1, 3$, and 5 predictor variables. A test statistic d below the relevant lower curve results in a rejection of the null hypothesis of zero serial correlation. If the test statistic is above the relevant upper curve, the null hypothesis is not rejected. If the test statistic is between the two curves, the test is indeterminate.

merator will tend to be larger. Therefore one does not reject the null hypothesis that the residuals are independent if the Durbin–Watson statistic is sufficiently large.

Figure 6.7 shows critical values for Durbin–Watson tests at the 5% level. These vary depending on the sample size and the number of predictor (x) variables, K. For simple linear regression, $K = 1$. For each value of K, Fig. 6.7 shows two curves. If the observed value of the test statistic falls below the lower curve, the null hypothesis is rejected, and one concludes that the residuals exhibit significant serial correlation. If the test statistic falls above the upper curve, the null hypothesis is not rejected, and one concludes that the residuals are not serially correlated. If the test statistic falls between the two relevant curves, the test is indeterminate.

Example 6.2. Examination of the Residuals from Example 6.1

A regression equation constructed using autocorrelated variables as predictand and predictors does not necessarily exhibit strongly autocorrelated residuals. Consider again the regression between Ithaca and Canandaigua minimum temperatures for January 1987 in Example 6.1. The lag-1 autocorrelations [Eq. (3.23)] for the Ithaca and Canandaigua minimum temperature data are 0.651 and 0.672, respectively. The residuals for this regression are plotted as a function of time in Fig. 6.8. A strong serial correlation for these residuals is not apparent, and their lag-1 autocorrelation as computed using Eq. (3.23) is only 0.191.

Fig. 6.8 Residuals from the regression, Eq. (6.19), plotted as a function of date. A strong serial correlation is not apparent, but the tendency for a negative slope indicates that the relationship between Ithaca and Canandaigua temperatures may be changing through the month.

Having computed the residuals for the Canandaigua–Ithaca minimum temperature regression, it is straightforward to compute the Durbin–Watson d [Eq. (6.20)]. In fact, the denominator is simply the SSE from the ANOVA Table 6.2, which is 341.622. The numerator in Eq. (6.20) must be computed from the residuals, and is 531.36. These yield $d = 1.55$. Referring to Fig. 6.7, the point at $n = 31$, $d = 1.55$ is well above the upper solid (for $K = 1$, since there is a single predictor variable) line, so the null hypothesis of uncorrelated residuals would not be rejected at the 5% level.

Finally, note that Fig. 6.8 reveals some additional information about the regression in Eq. (6.19). There is a perceptible trend for the residuals to decline through the month. Trends in the residuals, which may appear as slanted or curved bands in residual plots, indicate a different type of problem than the fanning in the residual plot in Fig. 6.4b. A linear regression for the 31 points in Fig. 6.8 yields a slope of $-0.162°$ F/day, with a standard error of $0.062°$ F/day. Thus this downward trend in Fig. 6.8 would be judged significant at the 5% level. Here it appears that the nature of the relationship may be changing through the month, perhaps as a consequence of the relationship between the annual cycles of temperatures at the two locations. As a practical matter, this trend in the residuals points out the need for one or more additional predictors (e.g., day of the month) in the regression equation. This could be easily achieved using multiple linear regression. ¤

6.2.7 Prediction Intervals

Frequently it is of interest to calculate confidence intervals around forecast values of the predictand. When it can be assumed that the residuals follow a Gaussian distribution, it is natural to approach this problem using the unbiasedness property

of the residuals [Eq. (6.8)], together with their estimated variance MSE $= s_\epsilon^2$. Using Gaussian probabilities (Table B.1), one expects a 95% confidence interval to be approximately bounded by $\hat{y} \pm 2s_\epsilon$.

The $\pm 2\sigma_\epsilon$ rule of thumb is often a quite good approximation to the width of a true 95% confidence interval, especially when the sample size is large. However, because both the sample mean of the predictand and the slope of the regression are subject to sampling variations, the prediction variance for data not used in the fitting of the regression is somewhat larger than the MSE. For a forecast of y using the predictand value x_0, the prediction variance is given by

$$s_{\hat{y}}^2 = s_\epsilon^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \tag{6.21}$$

That is, the prediction variance is proportional to the MSE, but is larger to the extent that the second and third terms inside the square brackets are appreciably larger than zero. The second term derives from the uncertainty in estimating the true mean of the predictand from a finite sample of size n [cf. Eq. (5.4)] and becomes much smaller than one for large sample sizes. The third term derives from the uncertainty in estimation of the slope [it is similar in form to Eq. (6.18b)], and indicates that predictions far removed from the center of the data used to fit the regression will be more uncertain than predictions made near the sample mean. However, even if the numerator in this third term is fairly large, the term itself will tend to be much smaller than one if a large data sample was used to construct the regression equation, since there are n nonnegative terms in the denominator.

Equation (6.21) is valid for simple linear regression. The corresponding equation for multiple regression is similar, but most conveniently expressed in matrix algebra notation (e.g., Draper and Smith, 1981; Neter *et al.*, 1985). However, in both cases the prediction variance is quite close to the MSE when a large sample size has been used to fit the regression.

6.2.8 Multiple Linear Regression

Multiple linear regression is the more general (and more common) situation of linear regression. As in the case of simple linear regression, there is still a single predictand, y, but in distinction there is more than one predictor (x) variable. The preceding treatment of simple linear regression was relatively lengthy, in part because most of what was presented generalizes readily to the case of multiple linear regression.

Let K denote the number of predictor variables. Simple linear regression then constitutes the special case of $K = 1$. The prediction equation [corresponding to Eq. (6.1)] becomes

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K. \tag{6.22}$$

Each of the K predictor variables has its own coefficient, analogous to the slope,

b, in Eq. (6.1). For notational convenience, the intercept (or regression "constant") is denoted as b_0 rather than as a as in Eq. (6.1). These regression coefficients are often called the *regression parameters*. In addition, the parameter b_0 is sometimes known as the *regression constant*.

Equation (6.2) for the residuals is still valid, if it is understood that the predicted value $\hat{y}$ is a function of a vector of predictors, x_k, $k = 1, \ldots, K$. If there are $K = 2$ predictor variables, the residual can still be visualized as a vertical distance. In this case, the regression function [Eq. (6.22)] is a surface rather than a line, and the residual corresponds geometrically to the distance above or below this surface along a line perpendicular to the (x_1, x_2) plane. The geometric situation is analogous for $K \geq 3$, but is not easily visualized.

The $K + 1$ parameters in Eq. (6.22) are found, as before, by minimizing the sum of squared residuals. This is achieved by simultaneously solving $K + 1$ equations analogous to Eqs. (6.5). This is most conveniently done using matrix algebra, the details of which can be found in standard regression texts (e.g., Draper and Smith, 1981; Neter *et al.*, 1985). The basics of the process are outlined in Example 9.2. In practice, the calculations are usually done using statistical software. The calculations are again summarized in an ANOVA table, of the form shown in Table 6.3. As before, SST is computed using Eq. (6.12), SSR is computed using Eq. (6.13a), and SSE is computed using the difference SST − SSR. The sample variance of the residuals is MSE = SSE/$(n - K - 1)$. The coefficient of determination is computed according to Eq. (6.16), although it is not the square of the Pearson correlation coefficient between the predictand and any of the predictor variables. The procedures sketched above for examination of residuals are applicable to multiple regression as well.

6.2.9 Derived predictor variables in multiple regression

Multiple regression opens up the possibility of an essentially unlimited number of potential predictor variables. An initial list of potential predictor variables can be multiplied manyfold by also considering mathematical transformations of these variables as potential predictors. Such "derived" predictors can be very useful in producing a good regression equation.

Table 6.3

Generic Analysis of Variance (ANOVA) Table for the Multiple Linear Regression[a]

Source	df	SS	MS	
Total	$n-1$	SST		
Regression	K	SSR	MSR = SSR / K	(F = MSR/MSE)
Residual	$n-K-1$	SSE	MSE = SSE / $(n-K-1) = s_\varepsilon^2$	

[a] Table 6.1 for simple linear regression can be viewed as a special case, with $K = 1$.

In some instances the form of the transformation may be suggested by the physics. In the absence of a strong physical rationale for particular variable transformations, the choice of a transformation or set of transformations may be made purely empirically, for example, by subjectively evaluating the general shape of the point cloud in a scatterplot. It may happen that the empirical choice of a transformation of the predictor variable in regression leads to a greater physical understanding, which is a highly desirable outcome in a research setting. This outcome would be less important for a purely forecasting problem, where the emphasis is on producing good forecasts rather than knowing precisely why the forecasts are good.

Transformations such as $x_2 = x_1^2$ or $x_2 = \sqrt{x_1}$, or any other power transformation of an available predictor can be regarded as another potential predictor. Similarly, trigonometric (sine, cosine, etc.), exponential or logarithmic functions, or combinations of these are useful in some situations. Another commonly used transformation is to a *binary*, or "dummy" variable. Binary variables take on one of two values (usually zero and 1, although the particular choices do not affect the forecast), depending on whether the variable being transformed is above or below a threshold or cutoff, c. That is, a binary variable x_2 could be constructed from another predictor x_1 according to the transformation

$$x_2 = \begin{cases} 1 & \text{if} \quad x_1 > c, \\ 0 & \text{if} \quad x_1 \leq c. \end{cases} \tag{6.23}$$

More than one binary predictor can be constructed from a single x_1 by choosing different values for the cutoff, c.

Even though transformed variables may be nonlinear functions of other variables, the overall framework is still known as multiple linear regression. Once a derived variable is defined, it is just another variable, regardless of how the transformation was made. More formally, the *linear* in multiple linear regression refers to the regression equation being linear in the parameters, b_k.

Example 6.3. A Multiple Regression with Derived Predictor Variables

Figure 6.9 is a scatterplot of a portion (1959–1988) of the famous Keeling monthly-averaged CO_2 concentration data from Mauna Loa in Hawaii. Representing the obvious time trend as a straight line yields the results shown in Table 6.4a. The regression line is also plotted (dashed) in Fig. 6.9. The results indicate a strong time trend, with standard error for the slope that is much smaller than the estimated slope. The intercept merely estimates the CO_2 concentration at $t = 0$, or December 1958, so the implied test for its difference from zero is of no interest. A literal interpretation of the MSE would indicate that a 95% confidence interval for measured CO_2 concentrations around the regression line would be about $\pm 2\sqrt{\text{MSE}} = 4.9$ ppm.

However, examination of a plot of the residuals versus time for this linear regression would reveal a "bowing" pattern, with a tendency for positive residuals

Fig. 6.9 A portion (1959–1988) of the Keeling monthly CO_2 concentration data (from Boden *et al.*, 1990), with linear (dashed) and quadratic (solid) least-squares fits.

at the beginning and end of the record, and with negative residuals being more common in the central part of the record. This can be discerned in Fig. 6.9 by noticing that most of the points fall above the dashed line at the beginning and end, and below the line toward the middle. A plot of the residuals versus the predicted values would also exhibit a bowing pattern, since there is a tendency for positive residuals for both high and low CO_2 concentrations, and negative residuals at intermediate concentrations.

This problem with the residuals can be alleviated (and the regression consequently improved) by fitting a quadratic curve to the time trend. To do this, a second predictor is added to the regression, and that predictor is simply the square of the time variable. That is, a multiple regression with $K = 2$ is fit using the predictors $x_1 = t$ and $x_2 = t^2$. Once defined, x_2 is just another predictor variable, taking on values between 0^2 and $360^2 = 129600$. The resulting least-squares quadratic curve is shown by the solid line in Fig. 6.9, and the corresponding regression statistics are summarized in Table 6.4b.

Of course, the SSTs in Tables 6.4a and 6.4b are the same since both pertain to the same predictand, the CO_2 concentrations. Both the coefficients $b_1 = 0.0501$ and $b_2 = 0.000136$ are substantially larger than their respective standard errors. The value of $b_0 = 312.9$ is again just the estimate of the CO_2 concentration at $t = 0$, and judging from the scatterplot, this intercept is a better estimate of its true value than was obtained from the linear regression. The data points are fairly evenly scattered around the quadratic trend line throughout the time period, so

Table 6.4

ANOVA Table and Regression Summary for Three Regression Fits to the 1959–1988 Portion of Keeling CO_2 Data in Fig. 6.9[a]

a. Linear Fit

Source	*df*	*SS*	*MS*	*F*
Total	356	39961.6		
Regression	1	37862.6	37862.6	6404
Residual	355	2099.0	5.913	

Variable	*Coefficient*	*Standard error*	*t-ratio*
Constant	312.9	0.2592	1207
t	0.0992	0.0012	90.0

b. Quadratic Fit

Source	*df*	*SS*	*MS*	*F*
Total	356	39961.6		
Regression	2	38483.9	19242.0	4601
Residual	354	1477.7	4.174	

Variable	*Coefficient*	*Standard error*	*t-ratio*
Constant	315.9	0.3269	966
t	0.0501	0.0042	12.0
t^2	0.000136	0.0000	12.2

c. Including Quadratic Trend for Annual Average and Harmonic Terms to Represent Annual Cycle

Source	*df*	*SS*	*MS*	*F*
Total	356	39961.6		
Regression	4	39783.9	9946.0	19696
Residual	352	177.7	0.5050	

Variable	*Coefficient*	*Standard error*	*t-ratio*
Constant	315.9	0.1137	2778
t	0.0501	0.0014	34.6
t^2	0.000137	0.0000	35.2
$\cos(2\pi t/12)$	−1.711	0.0530	−32.3
$\sin(2\pi t/12)$	2.089	0.0533	39.2

[a] The variable t (time) is a consecutive numbering of the months, with January 1959 = 1 and December 1988 = 360. There are n = 357 data points because Feburary–April 1964 are missing.

residual plots would exhibit the desired horizontal banding. Consequently, a 95% confidence interval of $\pm 2\sqrt{\text{MSE}} = 4.1$ ppm for CO_2 concentrations around the quadratic line would be applicable throughout the range of these data.

The quadratic function of time provides a reasonable approximation of the annual-average CO_2 concentration for the 30 years represented by the regression, although one can find periods of time where the point cloud wanders away from the curve. More importantly, however, a close inspection of the data points

in Fig. 6.9 reveals that they are not scattered randomly around the quadratic time trend. Rather, they execute a regular, nearly sinusoidal variation around the quadratic curve that is evidently an annual cycle. The resulting correlation in the residuals could be detected using the Durbin–Watson statistic, $d = 0.334$ (cf. Fig. 6.7). The CO_2 concentrations are lower in late summer and higher in late winter as a consequence of the annual cycle of photosynthetic carbon uptake by Northern Hemisphere plants, and respiratory carbon release from the dead plant parts. As will be shown in Chapter 8, this regular 12-month variation can be represented by introducing two more derived predictor variables into the equation: $x_3 = \cos(2\pi t/12)$ and $x_4 = \sin(2\pi t/12)$. Notice that both of these derived variables are functions only of the time variable t.

Table 6.4c indicates that, together with the linear and quadratic predictors included previously, these two *harmonic predictors* produce a very close fit to the data. The resulting prediction equation is

$$\begin{aligned}[CO_2] = {} & \underset{(0.1137)}{315.9} + \underset{(0.0014)}{0.0501}t + \underset{(.0000)}{0.000137}t^2 \\ & - \underset{(0.0530)}{1.711} \cos\left(\frac{2\pi t}{12}\right) + \underset{(0.0533)}{2.089} \sin\left(\frac{2\pi t}{12}\right),\end{aligned} \tag{6.24}$$

with all regression coefficients being much larger than their respective standard errors. The near equality of SST and SSR indicate that the predicted values are nearly coincident with the observed CO_2 concentrations [cf. Eqs. (6.12) and (6.13a)]. The resulting coefficient of determination is $R^2 = 39783.9/39961.6 = 99.56\%$, and the approximate 95% confidence interval implied by $\pm 2\sqrt{\text{MSE}}$ is only 1.4 ppm. A graph of Eq. (6.24) would wiggle up and down around the solid curve in Fig. 3.9, passing rather close to all the data points. ¤

6.3 Objective Forecasts—Without NWP

Construction of weather forecasts through purely statistical means, that is without the benefit of information from numerical (dynamical) weather prediction models, has come to be known as *classical* statistical forecasting. This name reflects the long history of the use of purely statistical forecasting methods, dating from before the availability of NWP information. Most classical forecast products are based on multiple regression equations. These statistical forecasts are *objective* in the sense that a particular set of inputs or predictors will always produce the same forecast for the predictand, once the forecast equation has been developed. However, many subjective decisions necessarily go into the development of the forecast equations.

The construction of a classical statistical forecasting procedure requires developmental data consisting of past values of the predictand to be forecast, and a matching collection of potential predictors whose values will be known prior to

the forecast time. A regression equation is first fit to this set of historical "training" data. This regression equation can then be used to forecast future values of the predictand on the basis of future observations of the predictor variables.

It is a characteristic of classical statistical weather forecasting that the time lag is built directly into the forecast equation through the time relationships between the predictors and the predictand. Therefore the time lag between observation of the predictors and the quantity to be predicted must be at least as long as the forecast projection. As a practical matter not all potential predictors become available simultaneously, and a finite amount of time is required for operational data processing and forecast dissemination. Predictors that are not available in time for the operational production of forecasts are of no value in a classical statistical forecasting procedure, regardless of the strength of their relationship with the predictand.

6.3.1 Stratification and Compositing

The physical and/or statistical relationships between predictors and predictand may change according to some known condition. Most commonly, the relationships can be different in different seasons. One approach to addressing seasonal differences is to include predictors that are functions of the day of the year. However, for some problems such differences can be better represented by stratifying the data according to the time of year, and producing separate forecast equations for perhaps each month or season. In other applications better forecasts might be achieved through stratification of the data according to meteorological or climatological criteria. For example, separate temperature forecast equations might be developed for snow-covered or open-ground conditions, or long-range forecast equations might be derived separately for El Niño versus non–El Niño years.

Conversely, it is sometimes useful to composite data sets in order to increase sample size, producing a single forecast equation for all of the combined cases. For example, data from nearby and climatologically similar stations might be combined to produce a forecast equation pertaining to any or all of them. This approach is especially useful for forecasting relatively rare events, for which fewer examples exist in the record of any one station than would be desirable. Individual station differences can still be incorporated into such pooled regression equations, for example, by including binary variables for individual stations as part of the potential predictor pool.

6.3.2 When the Predictand Is a Probability

One important advantage of statistical over (deterministic) dynamical forecasting methods is the capacity to produce probability forecasts. Inclusion of probability elements into the forecast format is advantageous because it provides an explicit expression of the inherent uncertainty or state of knowledge about the future

weather, and because probabilistic forecasts allow users to extract more value from them in making decisions (e.g., Thompson, 1962; Murphy, 1977; Krzysztofowicz, 1983). In a sense, almost any statistically based forecast procedure can produce probability information about a predictand, for example, by constructing a 95% confidence interval around a point regression forecast through application of the $\pm 2\sqrt{\text{MSE}}$ rule. More narrowly, however, probability forecasts are forecasts in which the predictand is a probability, rather than the value of a physical meteorological variable.

Most commonly, systems for producing probability forecasts are developed in a regression setting, by first transforming the predictand to a binary (or "dummy") variable taking on the values zero and one. That is, regression procedures are implemented after applying Eq. (6.23) to the predictand, y. In a sense, zero and one can be viewed as probabilities of the predictand not occurring or occurring, respectively, after the event has been observed.

There are two regression approaches to estimating predictand probabilities. The simplest, called *regression estimation of event probabilities* (REEP), involves only the exercise of multiple linear regression to derive a forecast equation for the binary predictand. The resulting predicted values are usually between zero and one, and it has been found through operational experience that these predicted values can be treated as specifications of probabilities for the event $\{Y = 1\}$. The main attraction to the use of REEP is that it is no more computationally demanding than any other linear regression problem, and has therefore been used extensively when computational resources have been limiting. It is somewhat unsatisfying theoretically, however, because it is possible to forecast probabilities that are either negative or greater than one. Even so, this causes little difficulty in an operational setting because forecast equations with many predictors rarely produce such nonsense probability estimates. When the problem does occur the forecast probability is usually near zero or one, and the operational forecast can be issued as such.

More theoretically satisfying is the use of a technique called *logistic regression.* Logistic regression also involves use of a binary predictand, but fits regression parameters to the nonlinear equation

$$\hat{y} = \frac{1}{1 + \exp[b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K]}. \tag{6.25}$$

The meaning of this equation is most easily understood for the single-predictor case of $K = 1$, for which it can be visualized geometrically as an S-shaped curve as a function of x_1. In the limits, $b_0 + b_1 x_1 \to +\infty$ means that the exponential function in the denominator becomes arbitrarily large so that the predicted value approaches zero. As $b_0 + b_1 x_1 \to -\infty$, the exponential function approaches zero and the predicted value approaches one. Thus it is guaranteed that logistic regression will produce properly bounded probability estimates. Depending on the parameters b_0 and b_1 the function rises gradually or abruptly from zero to one for

intermediate values of x_1. Note that the expression in the square brackets in the denominator of Eq. (6.25) is the same as for the usual multiple linear regression equation, Eq. (6.21), but that particular values for the b's will in general be very different even if the same predictor variables are chosen.

Equation (6.25) is a nonlinear function of the regression parameters. In general, analytic solutions as in Eq. (6.7) cannot be found for the parameters in Eq. (6.25). Rather, the parameters must be estimated using computationally intensive iterative techniques, and it is this requirement that has historically limited the use of logistic regression in statistical weather forecasting. However, many statistical computer packages include logistic regression as an option, and logistic regression is an increasingly practical option as computing resources become progressively more affordable.

Example 6.4. Comparison of REEP and Logistic Regression

Figure 6.10 compares the results of REEP (dashed) and logistic regression (solid) for some of the January 1987 data from Table A.1. The predictand is daily Ithaca precipitation, transformed to a binary variable using Eq. (6.23) with $c = 0$. That is, $y = 0$ if the precipitation is zero, and $y = 1$ otherwise. The predictor is the Ithaca minimum temperature for the same day. The parameters for both curves are chosen to minimize the sum of squared residuals. For the REEP (linear regression) equation, $b_0 = 0.208$ and $b_1 = 0.0212$. Thus, this equation specifies negative probability of precipitation if the temperature predictor is less than about $-9.8°$F, and specifies probability of precipitation greater than one if the mini-

Fig. 6.10 Comparison of regression probability forecasting using REEP (dashed line) and logistic regression (solid curve) using the January 1987 data set in Table A.1. Both functions were fit using least squares to the data shown by the dots. The binary predictand $y = 1$ if Ithaca precipitation is greater than zero and $y = 0$ otherwise.

mum temperature is greater than about 37.4° F. The parameters for the logistic regression are $b_0 = 1.846$ and $b_1 = -0.119$. The logistic regression curve produces similar probabilities through most of the temperature range, but these probabilities are constrained by the functional form of Eq. (6.25) to lie between zero and one even for extreme values of the predictor. ¤

6.3.3 Predictor Selection

There are almost always more potential predictors available than can be used in a regression, and finding good subsets of these in particular cases is more difficult than one might at first imagine. The process is definitely not as simple as adding members of the list of potential predictors to the regression until an apparently good relationship is achieved. Perhaps surprisingly, there are dangers associated with including too many predictor variables in a forecast equation.

Example 6.5. An Overfit Regression

To illustrate the dangers of too many predictors, Table 6.5 shows total winter snowfall at Ithaca (inches) for the seven winters beginning in 1980 through 1986, and four potential "predictors" arbitrarily taken from an almanac (Hoffman, 1988): the U.S. federal deficit (in billions of dollars), the number of personnel in the U.S. Air Force, the number of sheep in the United States (in thousands), and the average Scholastic Aptitude Test (SAT) scores of college-bound high-school students. Obviously these are nonsense predictors that bear no real relationship to the amount of snowfall at Ithaca.

Regardless of their lack of relevance, one can blindly offer these predictors to a computer regression package, and it will produce a regression equation. For reasons that will be made clear shortly, assume that the regression will be fit using

Table 6.5

A Small Data Set to Illustrate the Dangers of Overfitting[a]

Winter beginning	Ithaca snowfall (inches)	U.S. federal deficit ($\$ \times 10^9$)	U.S. Air Force personnel	U.S. sheep ($\times 10^3$)	Average SAT scores
1980	52.3	59.6	557,969	12,699	992
1981	64.9	57.9	570,302	12,947	994
1982	50.2	110.6	582,845	12,997	989
1983	74.2	196.4	592,044	12,140	963
1984	49.5	175.3	597,125	11,487	965
1985	64.7	211.9	601,515	10,443	977
1986	65.6	220.7	606,500	9,932	1001

[a] [Nonclimatological data were taken from Hoffman (1988).]

only the 6 years 1980 through 1985. That portion of available data used to produce the forecast equation is known as the *developmental sample*, *dependent sample*, or *training sample*. For the developmental sample of 1980–1985, the resulting equation is

$$\text{Snow} = 1161771 - 601.7(\text{year}) - 1.733(\text{deficit}) + 0.0567(\text{AF pers.}) - 0.3799(\text{sheep}) + 2.882(\text{SAT}),$$

where AF pers. represents U.S. Air Force personnel. The ANOVA table accompanying this equation indicated MSE = 0.0000, $R^2 = 100.00\%$, and $F = \infty$; that is, a perfect fit!

Figure 6.11 shows a plot of the regression-specified snowfall totals (line) and the observed data (circles). For the developmental portion of the record, the regression does, indeed, represent the data exactly, as indicated by the ANOVA statistics, even though it is obvious from the nature of the predictor variables that the specified relationship is not meaningful. In fact, any five predictors would have produced exactly the same perfect fit (although with different regression coefficients, b_k) to the six developmental data points. More generally, any $K = n - 1$ predictors will produce a perfect regression fit to any predictand for which there are n observations. This is easiest to see for the case of $n = 2$, where a straight line can be fit using any $K = 1$ predictor (simple linear regression), since a line

Fig. 6.11 "Forecasting" Ithaca winter snowfall using the data in Table 6.5. The number of predictors is one fewer than the number of observations of the predictand in the developmental data, yielding perfect correspondence between the values specified by the regression and the data for this portion of the record. The relationship falls apart completely when used with the 1986 data, which was not used in equation development. The regression equation has been grossly overfit.

can be found that will pass through any two points in the plane, and only an intercept and a slope are necessary to define a line. The problem, however, generalizes to any sample size.

This example illustrates an extreme case of "overfitting" the data. That is, so many predictors have been used that an excellent fit has been achieved on the dependent data, but the fitted relationship falls apart when used with *independent*, or *verification data*—data not used in the development of the equation. Here the data for 1986 have been reserved for a verification sample. Figure 6.11 indicates that the equation performs very poorly, indeed, producing a meaningless "forecast" for negative snowfall during 1986/87. Clearly, issuing forecasts equal to the climatological average total snowfall, or the snowfall for the previous winter, would yield better results that the equation produced above. Note that the problem of overfitting is *not* limited to cases where nonsense predictors are used in a forecast equation, and will be a problem when too many meaningful predictors are included as well. ¤

As ridiculous as it may seem, several important lessons can be drawn from Example 6.5:

1. Begin development of a regression equation by choosing only physically reasonable or meaningful potential predictors. If the predictand of interest is surface temperature, for example, then temperature-related predictors such as the 1000–700-mb thickness (reflecting the mean virtual temperature in the layer), the 700-mb relative humidity (perhaps as an index of clouds), or the climatological average temperature for the forecast date (as a representation of the annual cycle of temperature) could be sensible candidate predictors. Assuming that clouds will form only in saturated air, a binary variable based on the 700-mb relative humidity might also be expected to contribute meaningfully to the regression. Unpromising potential predictor variables might include upper-air heights or other meteorological observations at locations very far removed from the forecast location, the phase of the moon, or average stockmarket prices.

2. A tentative regression equation needs to be tested on a sample of data not involved in its development. One way to approach this important step is to simply reserve a portion (perhaps a quarter, a third, or half) of the available data as the independent verification set, and fit the regression using the remainder as the training set. The performance of the resulting regression equation will nearly always be better for the dependent than the independent data, since the coefficients have been chosen specifically to minimize the squared residuals in the developmental sample. A very large difference in performance between the dependent and independent samples would lead to the suspicion that the equation had been overfit.

3. One needs a reasonably large developmental sample if the resulting regression equation is to be *stable*. Stability is usually understood to mean that the regression coefficients are also applicable to independent (i.e., future) data, and have not resulted from an overfit regression. While the number of regression coefficients that can be estimated with reasonable accuracy increases as the sample

size increases, in forecasting practice it is often found that there is little to be gained from including more than about a dozen predictor variables in a final regression equation (Glahn, 1985). In this kind of forecasting application there are typically thousands of observations of the predictand in the developmental sample. Unfortunately, there is not a firm rule specifying a minimum ratio of data (number of observations of the predictand) to the number of predictor variables in the final equation. Rather, testing on an independent data set is relied upon in practice to ensure stability of the regression.

6.3.4 Screening Regression

Suppose the set of potential predictor variables for a particular regression problem could be assembled in a way that all physically relevant predictors were included, with exclusion of all irrelevant ones. This ideal can rarely, if ever, be achieved. Even if it could be, however, it would generally not be useful to include all of the potential predictors in a final regression equation. This is because the predictor variables are almost always mutually correlated, so that the full set of potential predictors contains redundant information. Table 3.2, for example, shows substantial correlations among the six variables in Table A.1. As a practical matter, then, one needs a method to choose among potential predictors, and of deciding how many of them are sufficient to produce a good prediction equation.

In the jargon of the atmospheric sciences, the problem of selecting a good set of predictors from a pool of potential predictors is called *screening regression*, since the predictors must be subjected to some kind of screening, or filtering procedure. The most commonly used screening procedure is known as *forward selection* or *stepwise regression* in the broader statistical literature.

Suppose that there are some number, M, of potential predictors. One begins the process of forward selection with the uninformative prediction equation $\hat{y} = b_0$. That is, only the intercept term is "in the equation," and the intercept is necessarily the sample mean of the predictand. On the first forward selection step, all M potential predictors are examined for the strength of their linear relationship to the predictand. In effect, all possible M simple linear regressions between the available predictors and the predictand are computed, and that predictor whose linear regression is best among all candidate predictors is chosen as x_1. At this stage of the screening procedure, then, the prediction equation is $\hat{y} = b_0 + b_1 x_1$. Note that the intercept b_0 will in general no longer be the average of the y values.

At the next stage of the forward selection, trial regressions are again constructed using all remaining $M - 1$ predictors. However, all of these trial regressions also contain the variable selected on the previous step as x_1. That is, given the particular x_1 chosen on the previous step, that predictor variable yielding the best regression $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ is chosen as x_2. This new x_2 will be recognized as "best" because it produces that regression equation with $K = 2$ predictors, also including the previously chosen x_1, having the highest R^2, the smallest MSE, and the largest F ratio.

Subsequent steps in the forward selection procedure follow this pattern exactly: At each step, that member of the potential predictor pool not yet in the regression is chosen that produces the best regression in conjunction with the $K - 1$ predictors chosen on previous steps. In general, when these regression equations are recomputed, the regression coefficients for the intercept and for the previously chosen predictors will change. These changes will occur because the predictors are usually correlated to a greater or lesser degree, so that the information about the predictand is "spread around" among the predictors differently as more predictors are added to the equation.

An alternative approach to screening regression is called *backward elimination*. The process of backward elimination is analogous but opposite to that of forward selection. Here the initial point is a regression containing all M potential predictors, $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Mx_M$. Usually this initial equation will be grossly overfit, containing many redundant and some possibly useless predictors. At each step of the backward elimination procedure, the least important predictor variable is removed from the regression equation. That variable will be the one whose coefficient is smallest, in an absolute value sense, with respect to its estimated standard error. In terms of the sample regression output tables presented above, the removed variable will exhibit the smallest (absolute) t ratio. As in forward selection, the regression coefficients for the remaining variables require recomputation if (as is usually the case) the predictors are mutually correlated.

There is no guarantee that forward selection and backward elimination will choose the same subset of the potential predictor pool for the final regression equation. Other variable selection procedures for multiple regression also exist, and these might select still different subsets. Although the possibility that a chosen variable selection procedure might not select the "right" set of predictor variables may be unsettling at first, as a practical matter this is rarely an important problem in the context of producing an equation for use as a forecast tool. Correlations among the predictor variables result in the situation that essentially the same information about the predictand can be extracted from different subsets of the potential predictors. Therefore, if the aim of the regression analysis is only to produce reasonably accurate forecasts of the predictand, the "black box" approach of empirically choosing a workable set of predictors is quite adequate. However, one should not be so complacent in a research setting, where one aim of a regression analysis could be to find specific predictor variables most directly responsible for the physical phenomenon associated with the predictand.

6.3.5 Stopping Rules

Both forward selection and backward elimination require a stopping criterion, or stopping rule. Without such a rule, forward selection would continue until all M candidate predictor variables were included in the regression equation, and backward elimination would continue until all predictors had been eliminated. At first blush, it might seem that finding the stopping point would be a simple matter of

evaluating the test statistics and their nominal p values as supplied by the computer regression package. Unfortunately, because of the way the predictors are selected, these implied hypothesis tests are not quantitatively applicable. At each step (either in selection or elimination) predictor variables are not chosen randomly for entry or removal. Rather, the best or worst, respectively, among the available choices is selected. While this may seem like a minor distinction, it can have very major consequences.

The problem is illustrated in Fig. 6.12, taken from the study of Neumann *et al.* (1977). The specific problem represented in this figure is the selection of exactly $K = 12$ predictor variables from pools of potential predictors of varying sizes, M, when there are $n = 127$ observations of the predictand. Ignoring the problem of nonrandom predictor selection would lead one to declare as significant any regression for which the F ratio in the ANOVA table is larger than the nominal value of 2.35. Naïvely, this value would correspond to the minimum F ratio necessary to reject the null hypothesis of no real relationship between the predictand and the 12 predictors, at the 1% level. The curve labeled "empirical F ratio" was arrived at using a resampling test (see Chapter 5), in which the same meteorological predictor variables were used in a forward selection procedure to predict 100 artificial data sets of $n = 127$ independent Gaussian random numbers each. This procedure simulates the null hypothesis that these same predictors bear no real relationship to the predictand.

Fig. 6.12 Comparison of the nominal and empirically (resampling) estimated critical ($p = 0.01$) F ratios for overall significance in a particular regression problem, as a function of the number of potential predictor variables, M. The sample size is $n = 127$, with $K = 12$ predictor variables to be included in each final regression equation. The "nominal" F ratio of 2.35 is applicable only for the case of $M = K$. When the forward selection procedure can choose from among more than K potential predictors, the true critical F ratio is substantially higher. The difference between the nominal and actual values widens as M increases. [After Neumann *et al.* (1977).]

Figure 6.12 indicates that the "nominal" value gives the correct answer only in the case of $K = M$, for which there is no ambiguity in the predictor selection since all the $M = 12$ potential predictors must be used to construct the $K = 12$ predictor equation. When the forward selection procedure has available some larger number $M > K$ potential predictor variables to choose from, the true critical F ratio is higher, and sometimes by a substantial amount. Even though none of the potential predictors in the resampling procedure bears any real relationship to the artificial (random) predictand, the forward selection procedure chooses those predictors exhibiting the highest chance correlations with the predictand, and these relationships result in apparently high F ratio statistics. Put another way, the p value associated with the nominal critical $F = 2.35$ is larger (less significant) than the true p value, by an amount that increases as more potential predictors are offered to the forward selection procedure. To emphasize the seriousness of the problem, the nominal F ratio in the situation of Fig. 6.12 for the very stringent $p = 0.0001$ level test is only about 3.7. The practical result of relying literally on the nominal critical F ratio is to allow more predictors into the final equation than are meaningful, with the danger that the regression will be overfit.

Unfortunately, the results in Fig. 6.12 apply only to the specific data set from which they were derived. In order to use this approach of estimating the true critical F-ratio using resampling methods it must be repeated for each regression to be fit, since the statistical relationships among the potential predictor variables will be different in different data sets. In practice, other less rigorous stopping criteria are usually employed. For example, one might stop adding predictors in a forward selection when none of the remaining predictors would reduce the R^2 by a specified amount, perhaps 0.05%.

The stopping criterion can also be based on the MSE. This is intuitively appealing since, as the standard deviation of the residuals around the regression function, $\sqrt{\text{MSE}}$ directly reflects the anticipated forecast accuracy of the regression. For example, if a regression equation were being developed to forecast surface temperature, little would be gained by adding more predictors if the MSE were already $0.01°\text{F}^2$, since this would indicate a $\pm 2\sigma$ (i.e., approximately 95%) confidence interval around the forecast value of about $\pm 2\sqrt{0.01°\text{F}^2} = 0.2°\text{F}$. So long as the number of predictors K is substantially less than the sample size n, adding more predictor variables (even meaningless ones) will decrease the MSE for the developmental sample. This concept is illustrated schematically in Fig. 6.13. Ideally the stopping criterion would be at the point where the MSE does not decline appreciably with the addition of more predictors, at perhaps $K = 12$ predictors in the hypothetical case shown in Fig. 6.13.

Figure 6.13 indicates that the MSE on an independent data set will be larger than that achieved for the developmental data. This result should not be surprising, since the least-squares fitting procedure operates by selecting precisely those parameter values that minimize the MSE for the developmental data. This underestimate of the operational MSE provided by the performance of a forecast equation on developmental data is an expression of what is sometimes called *artificial*

Fig. 6.13 Schematic illustration of the regression MSE as a function of the number of predictor variables in the equation, K, for developmental data (solid), and for an independent verification set (dashed). [After Glahn (1985).]

skill (Davis, 1976; Michaelson, 1987). The precise magnitude of the differences in MSE between developmental and independent data sets is not determinable solely from the regression output using the developmental data. That is, having seen only the developmental data, one cannot tell whether the minimum MSE for independent data will be at a similar point (at around $K = 12$ in Fig. 6.13), or whether the equation has been overfit and the minimum MSE for the independent data will be for a substantially smaller K. This situation is unfortunate, because the purpose of developing a forecast equation is to specify future, unknown values of the predictand using observations of the predictors that have yet to be observed.

Figure 6.13 also indicates that, for forecasting purposes, the exact stopping criterion is not generally critical as long as it is approximately right. Again, this is because the MSE tends to change relatively little through a range of K near the optimum, and for purposes of forecasting it is the minimization of the MSE rather than the specific identities of the predictors that is important. That is, in this case it is acceptable to use the forecasting regression equation as a "black box." By contrast, if the purpose of the regression analysis is scientific understanding, the specific identities of chosen predictor variables can be critically important, and the magnitudes of the resulting regression coefficients can lead to significant physical insight. In this case it is not reduction of prediction MSE per se that is desired, but rather that causal relationships between particular variables be suggested by the regression procedure.

Example 6.6. Equation Development Using Forward Selection

The concept of variable selection can be illustrated with the January 1987 temperature and precipitation data in Table A.1. As in Example 6.1 for simple linear

regression, the predictand is Canandaigua minimum temperature. The potential predictor pool consists of maximum and minimum temperatures at Ithaca, maximum temperature at Canandaigua, the logarithms of the precipitation amounts plus 0.01 in. (so the logarithm is defined for zero precipitation) for both locations, and the day of the month. The date predictor is included on the basis of the trend in the residuals apparent in Fig. 6.8. Note that this example is somewhat artificial with respect to statistical weather forecasting, since the predictors (other than the date) will not be known in advance of the time that the predictand (minimum temperature at Canandaigua) will be observed. However, this small data set serves perfectly well to illustrate the principles.

Figure 6.14 diagrams the process of choosing predictors using forward selection. The numbers in each table summarize the comparisons to be made at each step. For the first ($K = 1$) step, no predictors are yet in the equation, and all six potential predictors are under consideration. At this step the predictor producing the best simple linear regression is chosen, as indicated by the smallest MSE, and the largest R^2 and F ratio among the six. This best predictor is the Ithaca minimum temperature, so the tentative regression equation at this stage is exactly Eq. (6.19).

Having chosen the Ithaca minimum temperature in the first step, we find five potential predictors remaining, and these are listed in the $K = 2$ table. Of these five, the one producing the best predictions in an equation that also includes the Ithaca minimum temperature is chosen at this step. Summary statistics for these five possible two-predictor regressions are also shown in the $K = 2$ table. Of these,

K = 1

X	MSE	R^2	F
Date	51.1	36.3	16.5
Ith Max	33.8	57.9	39.9
Ith Min	**11.8**	**85.3**	**169***
Ith Ppt	65.0	19.0	6.80
CanMax	27.6	65.6	55.4
CanPpt	71.2	11.3	3.70

K = 2

X	MSE	R^2	F
Date	**9.2**	**88.9**	**112***
Ith Max	10.6	87.3	96.1
Ith Ppt	11.8	85.8	84.2
CanMax	10.0	88.0	103
CanPpt	10.5	87.3	96.3

K = 3

X	MSE	R^2	F
Ith Max	8.0	90.7	88.0
Ith Ppt	9.4	89.1	73.5
CanMax	**7.7**	**91.0**	**91.2***
CanPpt	8.6	90.0	80.9

K = 4

X	MSE	R^2	F
Ith Max	8.0	91.0	65.9
Ith Ppt	8.0	91.1	66.6
CanPpt	**7.7**	**91.4**	**69.0***

K = 5

X	MSE	R^2	F
Ith Max	8.0	91.4	53.4
Ith Ppt	**6.8**	**92.7**	**63.4***

Fig. 6.14 Diagram of the forward selection procedure for development of a regression equation for Canandaigua minimum temperature using as potential predictors the remaining variables in the data set from Table A.1, plus the date. At each step the variable is chosen (bold, starred) whose addition would produce the largest decrease in MSE or, equivalently, the largest increase in R^2 or F. At the final ($K = 6$) stage, only IthMax remains to be chosen, and its inclusion would produce MSE = 6.8, $R^2 = 93.0\%$, and $F = 52.8$. (*Key*: Can—Canandaigua; Ith—Ithaca; Max—maximum; Min—minimum; Ppt—precipitation).

the equation including Ithaca minimum temperature and the date as the two predictors is clearly best, producing MSE = 9.2° F^2 for the dependent data.

With these two predictors now in the equation, there are only four potential predictors left at the $K = 3$ stage. Of these, the Canandaigua maximum temperature produces the best predictions in conjunction with the two predictors already in the equation, yielding MSE = 7.7° F^2 on the dependent data. Similarly, the best predictor at the $K = 4$ stage is Canandaigua precipitation, and the better predictor at the $K = 5$ stage is Ithaca precipitation. For $K = 6$ (all predictors in the equation) the MSE for the dependent data is 6.8° F^2, with $R^2 = 93.0\%$.

Where this selection process would stop depends on the stopping rule employed. If the rule specifies stopping at a leveling off of the dependent sample MSE, probably the $K = 3$ predictor equation using Ithaca minimum temperature, the date, and Canandaigua maximum temperature as predictors would be chosen. This is because the decrease in MSE achieved by adding Canandaigua precipitation at the $K = 4$ stage is only 0.035° F^2. On the other hand, if the rule were to add predictors until the R^2 increases less than 0.5%, the $K = 5$ regression would be chosen. The $K = 5$ equation contains all of the original six potential predictors except the Ithaca maximum temperature. ¤

6.3.6 Cross-Validation

Usually regression equations to be used for weather forecasting are tested on a sample of independent data that has been held back during the development of the forecast equation. In this way, once the number K and specific identities of the predictors have been fixed, the distances between the solid and dashed MSE lines in Fig. 6.13 can be estimated directly from the reserved data. If the deterioration in forecast precision (i.e., the increase in MSE) is judged to be acceptable, the equation can be used operationally.

This procedure is actually a special case of a technique known as *cross-validation* (Efron 1982; Efron and Gong, 1983; Michaelson, 1987). Cross-validation is a resampling technique whose mechanics are analogous to the bootstrap and permutation tests (Chapter 5). In cross-validation the available data are repeatedly divided into developmental and verification data subsets. Simply evaluating the performance of a forecast equation development algorithm on part of the data, while using the remainder of the data as the independent set as described in the previous paragraph, constitutes a single exercise of the cross-validation procedure.

Very often, cross-validation is carried out using developmental data sets of size $n - 1$, and verification data "sets" containing the remaining single observation of the predictand. In this case there are clearly n distinct partitions of the data. The regression model is recalculated for each of these partitions. The result is n similar forecast equations, each computed without one of the observations of the predictand. The cross-validation estimate of the prediction MSE is then computed by forecasting each omitted observation using that equation developed from the re-

maining $n - 1$ data values, computing the squared difference between the prediction and predictand for each of these equations, and averaging the n squared differences. Thus, the cross-validation procedure uses all n observations of the predictand to estimate the prediction MSE in a way that allows each observation to be treated, one at a time, as independent data. Note that cross-validation can also be carried out for any number m of withheld data points, and developmental data sets of size $n - m$. In this more general case, all $(n!)/[(m!)(n - m)!]$ possible partitions of the full data set could be employed.

In the context of the development of regression equations, the cross-validation procedure can also be applied for several stopping rules (implicitly, different numbers of predictors, K), to choose that rule yielding least prediction error on data not used in equation development. As a practical matter this requires very substantial computing resources for large problems, and has only recently become a feasible option. In any case, however, the ultimate "product" equation, to be used for operational forecasts, would be fit using all the data after one is satisfied with the cross-validation results.

Example 6.7. Protecting against Overfitting Using Cross-Validation

Having used all the available developmental data to fit the regressions in Example 6.6, what can be done to ensure that these prediction equations have not been overfit? Fundamentally, what is desired is a measure of how the regressions will perform when used on data not used to fit their parameters. Cross-validation is an especially appropriate tool for this purpose in the present example, because the small ($n = 31$) sample would be inadequate if a substantial portion of it had to be reserved for a validation sample.

Figure 6.15 compares the regression MSE for the six regression equations obtained with the forward selection procedure outlined in Fig. 6.14. This figure shows real results in the same form as the idealization of Fig. 6.13. The solid line indicates the MSE achieved on the developmental sample, corresponding to the bold MSE values in Fig. 6.14. Since each regression chooses precisely those coefficients minimizing this quantity on the developmental data, the MSE is expected to be higher when the equations are applied to independent data. An estimate of how much higher is given by the average MSE from the cross-validation samples (dashed line). Each cross-validation point represents the average of the 31 squared differences between an observed value of the predictand, and the forecast of that value produced by regression equations fit to all the data except that one point. That is, having selected the predictors to be used for each value of K in Fig. 6.14, each regression is recomputed using these predictors 31 times: once each for all the possible $n = 30$ subsets of the original data. The resulting equations are used to predict the one data point not used in their development.

The dashed line indicates that these regressions should produce prediction MSE on independent data of around $9.5°\text{F}^2$ (yielding $\pm 2\sigma$ confidence limits of $\pm 6.2°\text{F}$) for any of the values of K between 3 and 5. Exactly which of the predic-

Fig. 6.15 Plot of residual mean-squared error as a function of the number of regression predictors specifying Canandaigua minimum temperature, using the January 1987 data in Appendix A. The solid line shows MSE for developmental data (bold MSE numbers in Fig. 6.13). The dashed line shows MSE achievable on independent data, as estimated through cross-validation. Circle and "+" show the respective MSE values for the $K = 5$ predictor regression equation obtained using backward elimination. This plot is a real-data example corresponding to the idealization in Fig. 6.12.

tors are used in these candidate forecast equations is not as important as might have been anticipated, since the predictors exhibit substantial correlation. To underscore this point, the circle and "+" in Fig. 6.15 show respectively the developmental and cross-validation MSE for the $K = 5$ regression equation produced by the backward elimination predictor selection procedure. This equation differs from the $K = 5$ forward selection equation by including the Ithaca maximum temperature instead of the Canandaigua maximum temperature. Since the maximum temperatures at the two locations are strongly correlated (cf. Table 3.2 or Fig. 3.18), it makes little difference which of the two is included. The $K = 6$ cross-validation estimate of the MSE is $10.4°\,F^2$, which indicates that if both Ithaca and Canandaigua maximum temperatures are included as predictors, an overfit prediction equation results. ¤

Example 6.8. A Set of Classical Statistical Forecast Equations

The flavor of a "classical" statistical forecast method can be appreciated by looking at the NHC-67 procedure for forecasting hurricane movement (Miller *et al.*, 1968). This relatively simple set of regression equations was used as part of the operational suite of forecast models at the U.S. National Hurricane Center until 1988 (Sheets, 1990). Since hurricane movement is a vector quantity, each forecast consists of two equations: one for northward movement and one for westward

movement. The two-dimensional forecast displacement is then computed as the vector sum of the northward and westward displacements.

The predictands were stratified according to two geographical regions: north and south of 27.5° N latitude. That is, separate forecast equations were developed to predict storms on either side of this latitude. Separate forecast equations were also developed for "slow" versus "fast" storms. The choices of these two stratifications were made subjectively, on the basis of the experience of the developers of the forecast procedure regarding the behavior of Atlantic hurricanes. Separate equations are also needed for each forecast projection [0–12, 12–24, 24–36, 36–48, and 48–72 hours (h)], yielding a total of 2 (displacement directions) × 2 (regions) × 2 (speeds) × 5 (projections) = 40 separate regression equations in the NHC-67 package.

The available developmental data set consisted of 236 "northern" cases and 224 "southern" cases. Candidate predictor variables were derived primarily from 1000-, 700-, and 500-mb heights at each of 120 gridpoints in a 5° × 5° coordinate system that follows the storm. Predictors derived from these 3 × 120 = 360 geopotential height predictors, including 24-h height changes at each level, geostrophic winds, thermal winds, and Laplacians of the heights, were also included as candidate predictors. Additionally, two persistence predictors, observed northward and westward storm displacements in the previous 12 h were included.

With vastly more potential predictors than observations, some screening procedure is clearly required. Here forward selection was used, with the (subjectively determined) stopping rule that no more than 15 predictors would be in any equation, and new predictors would be included that increased the regression R^2 by at least 1%. This second criterion was apparently sometimes relaxed for regressions with few predictors.

Table 6.6 shows the results for the 0–12-h westward displacement of "slow" southern storms in NHC-67. The five predictors are shown in the order selected by the forward selection procedure, together with the R^2 value achieved on the developmental data at each step. The coefficients are those for the final ($K = 5$) equation. The most important single predictor was the persistence variable (P_x), reflecting the tendency of hurricanes to change speed and direction fairly slowly. The 500-mb height north and west of the storm (Z_{37}) corresponds physically to the "steering" effects of midtropospheric flow on hurricane movement. Its coefficient is positive, indicating a tendency for westward storm displacement given relatively high heights to the northwest, and slower or eastward (negative westward) displacement of storms located southwest of 500-mb troughs. The final two predictors appear to improve the regression only marginally—the predictor Z_3 increases the R^2 by less than 1%—and one suspects that the $K = 3$ predictor model might have done as well on independent data. Remarks in Neumann *et al.* (1977) concerning the fitting of the similar NHC-72 regressions, in relation to Fig. 6.12, are also consistent with the idea that the equation represented in Table 6.6 may have been overfit. ¤

Table 6.6

Regression Results for the NHC-67 Hurricane Model for the 0–12-h Westward Displacement of "Slow" Southern Zone Storms[a]

Predictor	Coefficient	Cumulative R^2
Intercept	−2709.5	—
P_X	0.8155	79.8%
Z_{37}	0.5766	83.7%
P_Y	−0.2439	84.8%
Z_3	−0.1082	85.6%
P_{51}	−0.3359	86.7%

[a] Indicating the order in which the predictors were selected and the resulting R^2 at each step. The meanings of the symbols for the predictors are P_X=westward displacement in the previous 12 h, Z_{37}=500-mb height at the point 10° north and 5° west of the storm, P_Y=northward displacement in the previous 12 h, Z_3=500-mb height at the point 20° north and 20° west of the storm, and P_{51}=1000-mb height at the point 5° north and 5° west of the storm. Distances are in nautical miles, and heights are in meters. [From Miller *et al.* (1968), reprinted with permission of the American Meteorological Society.]

6.3.7 Analog Forecasting

Not all objective statistical forecast procedures are based on regression. A number of such approaches were in use before the advent of reasonably accurate NWP forecasts for the 12–48-h projection range. Some of these objective forecast methods are briefly described in Glahn (1985) and Wadsworth (1951). One such procedure, analog forecasting, is still in use in purely classical form for long-range (seasonal) forecasting (Wagner, 1989), although the climatological database is too short for analog forecasts to be competitive in ordinary short-range weather forecasting (Gutzler and Shukla, 1984; Lorenz, 1969).

The idea underlying analog forecasting is to search archives of climatological synoptic data for maps closely resembling current observations, and assume that the future evolution of the atmosphere will be similar to the flows that followed the historical analogs. The method has intuitive appeal in that, unlike NWP computer models, the atmosphere is a perfect model of itself. As a practical matter, however, analog forecasting is limited by the fact that the atmosphere apparently does not ever exactly repeat itself. Therefore, matches between currently observed and historical meteorological fields (in the context of weather forecasting a *field* is a spatial array of meteorological quantities) can be only approximate. The problem is further complicated since it is not just a single field (say, hemispheric 500-mb heights), but many (in theory all relevant) fields that must be matched. For seasonal forecasting, not only such atmospheric fields as heights and temperatures are relevant, but also more slowly varying fields such as sea–surface temperature

and ice extent need to be matched. Obviously, the prospect of a good analog decreases with the number of features that must be matched.

Statistical ideas come into analog forecasting in the process of searching for analogs and ranking historical cases for closeness to the current observations. While the number of ways that two fields can be compared is essentially limitless, Toth (1991) lists some of the possibilities. A currently operational analog prediction system (Livezey and Barnston, 1988) uses a comparison based on principal-component analysis (see Chapter 9). It is perhaps easiest to think in terms of the correlation between the two fields to be compared. For computational convenience it is usual for meteorological fields to be approximated as a set of values at a collection of gridpoints. If two fields have been interpolated onto the same grid of n points, it is straightforward to compute the Pearson correlation [Eq. (3.17)] between them. The resulting correlation coefficient is sometimes called a *pattern correlation*, since pairs of maps with the similar patterns will yield high correlations.

Besides identifying analogs, statistics such as the pattern correlation allow ranking of analog cases with respect to the closeness of their match to the observed conditions. When more than one analog can be identified, the forecast can be constructed as an average of the outcomes that followed the analog cases, perhaps weighted according to the closeness of each analogy. Remarkably, it has been found that *antianalogies* (historical maps exhibiting the same patterns of anomalies but of opposite sign to the current map to be matched) can also be used to expand the number of available analogs, provided that the forecast is constructed from the reverse of the outcome following the antianalog (Harnack *et al.*, 1985).

6.4 Objective Forecasts—With NWP

Pure classical statistical weather forecasts for projections in the range of a few days are generally no longer employed, since current dynamical NWP models allow more accurate forecasts at this time scale. However, two newer types of statistical weather forecasting have arisen that can improve on aspects of the NWP forecasts. Both of these methods use large multiple regression equations in a way that is analogous to the classical approach, so that many of the same technical considerations pertaining to equation fitting apply. The differences between these two approaches and classical statistical forecasting lie in the range of available predictor variables. In addition to predictors such as the date or the climatological average values of a particular meteorological element, predictor variables taken from the output of the NWP models are also used.

There are three basic reasons why statistical reinterpretation of dynamical NWP output is useful for practical weather forecasting:

- There are important differences between the real world and its representation in NWP models. Figure 6.16 illustrates some of these. The NWP models necessar-

Fig. 6.16 Schematic illustration of the differences between the real world (a) and the world (b) as represented by numerical weather prediction models. [From Karl *et al.* (1989).]

ily simplify and homogenize surface conditions, by representing the world as an array of gridpoints, to which the NWP output pertains. As implied by Fig. 6.16, small-scale effects (e.g., of topography or small bodies of water) important to local weather may not be included in the NWP model. Also, locations and variables for which forecasts are desired may not be represented explicitly by the NWP model. However, statistical relationships can be developed between the information provided by the NWP models and desired forecast quantities to help alleviate these problems.

- The NWP models are not perfect, and their forecasts are subject to errors. To the extent that these errors are systematic, statistical forecasts based on the NWP information can compensate and correct forecast biases.
- The NWP models are deterministic. That is, even though the future state of the weather is inherently uncertain, these models are capable of producing only a single forecast for any meteorological element, given a set of initial model conditions. Using NWP information in conjunction with statistical methods allows quantification and expression of the uncertainty associated with different forecast situations. In particular, it is possible to derive probability forecasts, using REEP or logistic regression, using predictors taken from the deterministic NWP models.

6.4.1 Perfect Prog Forecasts

The first statistical approach to be developed for taking advantage of the deterministic dynamical forecasts from the NWP models (Klein *et al.*, 1959) is called *perfect prog*, which is short for "perfect prognosis." As the term implies, the perfect-prog technique makes no attempt to correct for possible NWP model errors or biases, but takes their forecasts for future atmospheric variables at face value—assuming them to be perfect.

Development of perfect prog regression equations is similar to the development of classical regression equations, in that observed predictors are used to specify

observed predictands. That is, only historical climatological data are used in the development of a perfect-prog forecasting equation. The primary difference between development of classical and perfect-prog equations is in the time lag. Classical equations incorporate the forecast time lag by relating predictors available before the forecast must be issued (say, "today") to values of the predictand to be observed at some later time (say, "tomorrow"). Perfect-prog equations do not incorporate any time lag. Rather, simultaneous values of predictors and predictands are used to fit the regression equations. That is, the equations specifying "tomorrow's" predictand are developed using "tomorrow's" predictor values.

At first, it might seem that this would not be a productive approach to forecasting. While tomorrow's 1000–850-mb thickness may be an excellent predictor for tomorrow's temperature, tomorrow's thickness will not be known until tomorrow. However, in implementing the perfect-prog approach, it is the NWP forecasts of the predictors (e.g., the forecast for tomorrow's thickness) that are substituted into the regression equation as predictor values. Therefore, the forecast time lag in the perfect-prog approach is contained entirely in the NWP model. Of course, this means that quantities not forecast by the NWP model cannot be included as potential predictors unless they will be known today. If the NWP forecasts for tomorrow's predictors really are perfect, the perfect-prog regression equations should provide very good forecasts.

6.4.2 Model Output Statistics (MOS) Forecasts

The MOS approach (Glahn and Lowry, 1972) is the second, and usually preferred, approach to incorporating NWP forecast information into statistical weather forecasts. The preference for the MOS approach derives from its capacity to include directly in the regression equations the influences of specific characteristics of different NWP models at different projections into the future.

Both the MOS and perfect-prog approaches use quantities from NWP output as predictor variables; however, these two approaches use the information differently. The perfect-prog approach only uses the NWP forecast predictors when making forecasts, but the MOS approach uses these predictors in both the development and implementation of the statistical equations. Think again in terms of "today" as the time at which the forecast must be made and "tomorrow" as the time to which the forecast pertains. The MOS regression equations are developed for tomorrow's predictand using NWP forecasts for tomorrow's values of the predictors. The true values of tomorrow's predictors are unknown today, but the NWP forecasts for those quantities are known today. For example, in the MOS approach one important predictor for tomorrow's temperature could be tomorrow's 1000–850-mb thickness *as forecast today* by a particular NWP model. Therefore, to develop MOS forecast equations it is necessary to have a developmental data set composed of historical records of the predictand, together with archived records of the forecasts produced by the NWP model for the same days on which the predictand was observed.

In common with the perfect-prog approach, the time lag in MOS forecasts is incorporated through the NWP forecast. Unlike perfect prog, the implementation of a MOS forecast equation is completely consistent with its development. That is, in both development and implementation, the MOS statistical forecast for tomorrow's predictand is made using the NWP dynamical forecast for tomorrow's predictors, which are available today. Also unlike the perfect-prog approach, separate MOS forecast equations must be developed for different forecast projections. This is because the characteristics of the NWP forecasts are different at different projections, producing, for example, different statistical relationships between observed temperature and forecast thicknesses for 24 h versus 48 h in the future.

6.4.3 Comparisons between the Classical, Perfect Prog, and MOS Approaches

The classical, perfect prog, and MOS approaches are all based on multiple regression, exploiting correlations between the predictand and available predictors. In the classical approach it is the correlations between today's values of the predictors and tomorrow's predictand that forms the basis of the forecast. For the perfect-prog approach it is the simultaneous correlations between today's values of both predictand and predictors that are the statistical basis of the prediction equations. In the case of MOS forecasts, the prediction equations are constructed on the basis of correlations between NWP forecasts as predictor variables, and the observed value of tomorrow's predictand.

These distinctions be expressed mathematically, as follows. In the classical approach, the forecast predictand at some future time, t, is expressed in the regression function f_C using a vector of (i.e., multiple) predictor variables, $\mathbf{x}_0$, according to

$$\hat{y}_t = f_C(\mathbf{x}_0). \tag{6.26}$$

The subscript 0 on the predictors indicates that they pertain to values observed at or before the time that the forecast must be formulated, which is earlier than the time t to which the forecast pertains. This equation emphasizes that the forecast time lag is built into the regression equation. This equation pertains both to the development and implementation of a classical statistical forecast equation.

By contrast, the perfect-prog (PP) approach operates differently for development versus implementation of the forecast equation, and these two phases can be expressed as

$$\hat{y}_0 = f_{PP}(\mathbf{x}_0) \qquad \text{in development,} \tag{6.27a}$$

and

$$\hat{y}_t = f_{PP}(\mathbf{x}_t) \qquad \text{in implementation,} \tag{6.27b}$$

respectively. The perfect-prog regression function, f_{PP}, is the same in both cases, but it is developed entirely with observed data having no time lag with respect to the predictand. In implementation it operates on forecast values of the predictors for the future time t, as obtained from the NWP model.

Finally, the MOS approach uses the same equation in development and implementation,

$$\hat{y}_t = f_{\mathrm{MOS}}(\mathbf{x}_t). \tag{6.28}$$

It is derived using the NWP forecast predictors $\mathbf{x}_t$, pertaining to the future time t, and is implemented in the same way. In common with the perfect-prog approach, the time lag is carried by the NWP forecast, not the regression equation.

Figure 6.17 illustrates these relationships between the classical, perfect prog, and MOS approaches to statistical forecasting. Development of the forecast equations is indicated by the solid arrows. For each forecasting procedure, one solid arrow points from a source of potential predictor (x) variables, and one points from a source of the developmental observations of the predictand (y). The classical forecast equations are produced using predictors available today together with tomorrow's value of the predictand, with both predictors and predictand being observed values from the real atmosphere. The perfect-prog equations are also fitted using observed values of the predictors and predictands from the real atmosphere, but without a time lag: The predictors and predictand are all observed on the same day. The MOS forecast equations are developed using predictors that are

Fig. 6.17 Schematic comparison of classical, perfect-prog, and MOS approaches to statistical weather forecasting. Solid arrows indicate information sources for development of each type of forecast equation, and dashed arrows indicate the information flow when each of the types of statistical forecasting systems are implemented.

forecasts from the NWP model (pertaining to tomorrow but available today) and observed predictands from the real atmosphere.

The dashed arrows in Fig. 6.17 indicate the flow of information when each of the three types of equations is used for forecasting. Since the object in each case is to forecast a future value of the predictand, dashed arrows from each of the three kinds of forecast equations point to tomorrow's predictand in the real atmosphere. For both the classical and MOS approaches, the implementation is consistent with the development. That is, the same sets of predictors and predictands are used in development and forecasting, so that the solid and dashed arrows are parallel. In the classical approach, today's real-atmosphere predictors are put through the regression equation to forecast tomorrow's real-atmosphere predictand. In the MOS approach, today's NWP model forecast for tomorrow contains the predictors used in the regression equation to forecast tomorrow's predictand. For the perfect-prog approach, the solid and dashed arrows are not parallel, illustrating that different routes are taken in development and implementation of the regression equations. The perfect-prog equations are developed with predictors observed simultaneously with the predictand, but when the equations are implemented the NWP forecasts for those predictors are substituted for their observed counterparts in the real atmosphere.

Given the comparatively high accuracy of dynamical NWP forecasts for weather forecasts, classical statistical forecasting is not competitive for forecasts in the range of a few days. Since the perfect-prog and MOS approaches both draw on this same source of information, it is worthwhile to compare their advantages and disadvantages.

There is nearly always a large developmental sample for perfect-prog equations, since these are developed using only historical climatological data. This is an advantage over the MOS approach, since fitting MOS equations requires an archived record of forecasts from the same NWP model that will ultimately be used to drive the MOS equations. Typically, several years of archived NWP forecasts are required to develop a stable set of MOS forecast equations. This requirement can constitute a substantial limitation, because the dynamical NWP models are not static. Rather, these models regularly undergo changes aimed at improving their performance. Minor changes in the NWP model leading to reductions in the magnitudes of random errors will not substantially degrade the performance of a set of MOS equations (e.g., Erickson *et al.*, 1991). However, modifications to the NWP model that change—even substantially reduce—systematic errors in the NWP model will require redevelopment of accompanying MOS forecast equations. Since it is a change in the NWP model that will have necessitated the redevelopment of a set of MOS forecast equations, it is likely that a sufficiently long developmental sample of predictors from the improved NWP model will not be immediately available. By contrast, since the perfect-prog equations are developed without any NWP information, changes in the NWP models should not require changes in the perfect-prog regression equations. Furthermore, improving

either the random or systematic error characteristics of the NWP model should improve the statistical forecasts produced by a perfect-prog equation.

Similarly, the same perfect-prog regression equations can in principle be used with any NWP model, or for any forecast projection provided by a single NWP model. Since the MOS equations are "tuned" to the particular error characteristics of the model for which they were developed, different MOS equations will in general be required for use with different dynamical NWP models. Analogously, since the error characteristics of the NWP model change with increasing projection, different MOS equations are required for forecasts of the same atmospheric variable for different projection times into the future. Note, however, that potential predictors for a perfect-prog equation must be variables that are well predicted by the NWP model with which they will be used. It may be possible to find an atmospheric predictor variable that relates closely to a predictand of interest, but which is badly forecast by a particular NWP model. Such a variable might well be selected for inclusion in a perfect-prog equation on the basis of the relationship of its observed values to the predictand, but would be ignored in the development of a MOS equation if the poor NWP forecast of that predictor bore little relationship to the predictand.

The MOS approach to statistical forecasting has one dominant advantage over the perfect-prog approach that makes MOS the method of choice when practical. This is that systematic errors exhibited by the dynamical NWP model are accounted for in the process of developing the MOS equations. Since the perfect-prog equations are developed without reference to the characteristics of the NWP model, they cannot account for or correct for any type of NWP forecast error. The MOS development procedure allows compensation for these systematic errors when forecasts are constructed. Systematic errors include such problems as progressive cooling or warming biases in the NWP model with increasing forecast projection, a tendency for modeled synoptic features to move too slowly or too quickly in the NWP model, and even the unavoidable decrease in forecast accuracy at increasing projections.

The compensation for systematic errors in a dynamical NWP model that is accomplished by MOS forecast equations is easiest to see in relation to a simple bias in an important predictor. Figure 6.18 illustrates a hypothetical case, where surface temperature is to be forecast using the 1000–850-mb thickness. The "x's" in the figure represent the (unlagged, or simultaneous) relationship of a set of observed thicknesses with observed temperatures, and the circles represent the relationship between NWP-forecast thicknesses with the same temperature data. As drawn, the hypothetical NWP model tends to forecast thicknesses that are too large by about 15 m.

The "observed" thicknesses (x's) in Fig. 6.18 appear to forecast the simultaneously observed surface temperatures reasonably well, yielding an apparently good perfect-prog regression equation (dashed line). The relationship between forecast thickness and observed temperature represented by the MOS equation

Fig. 6.18 Illustration of the capacity of a MOS equation to correct for systematic bias in a hypothetical NWP model. The "x's" represent observed, and the circles represent NWP-forecast 1000–850-mb thicknesses (Th), in relation to hypothetical surface temperatures. The bias in the NWP model is such that the forecast thicknesses are too large by about 15 m, on average. The MOS equation (solid line), is calibrated for this bias, and produces a reasonable temperature forecast (lower horizontal arrow) when the NWP forecast thickness is 1300 m. The perfect-prog equation (dashed line) incorporates no information regarding the performance of the NWP model, and produces a surface temperature forecast (upper horizontal arrow) that is too warm as a consequence of the thickness bias.

(solid line) is substantially different, because it includes the tendency for this NWP model to systematically produce thicknesses that are too large. If the NWP model produces a thickness forecast of 1300 m (vertical arrows), the MOS equation allows for bias in the thickness forecast and produces a reasonable temperature forecast of about 30° F (lower horizontal arrow). Loosely speaking, the MOS "knows" that when this NWP model forecasts 1300 m, a more reasonable expectation for the true future thickness is closer to 1285 m, which in the climatological data (x's) corresponds to a temperature of about 30° F. The perfect-prog equation, on the other hand, operates under the assumption that the NWP model will forecast the future thickness perfectly. It therefore yields a temperature forecast that is too warm (upper horizontal arrow) when supplied with a thickness forecast that is too large.

A more subtle systematic error exhibited by all NWP models is the degradation of forecast accuracy at increasing projections. The MOS approach to statistical

Fig. 6.19 Illustration of the capacity of a MOS equation to account for the systematic tendency of NWP forecasts to become less accurate at longer projections. The points in these panels are simulated thickness forecasts, constructed from the "x's" in Fig. 6.18 by adding random errors to the thickness values. As the forecast accuracy degrades at longer projections, the perfect-prog equation (dashed lines, reproduced from Fig. 6.18) is increasingly overconfident, and tends to forecast extreme temperatures too frequently. At longer projections (b) the MOS equations increasingly provide forecasts near the average temperature (30.8° F in this example).

forecasting accounts for this type of systematic error as well. The situation is illustrated in Fig. 6.19, which is based on the hypothetical "observed data" in Fig. 6.18. Panels a and b in Figure 6.19 simulate the relationships between forecast thicknesses from an unbiased NWP model at 24- and 48-h projections and the surface temperature, and have been constructed by adding random errors to the "observed" thickness values (x's) in Fig. 6.18. These random errors exhibit $\sqrt{\text{MSE}} = 20$ m for the 24-h projection and $\sqrt{\text{MSE}} = 30$ m at the 48-h projection. The increased scatter of points for the simulated 48-h projection illustrates that the regression relationship is weaker when the NWP model is less accurate.

The MOS equations (solid lines) fit to the two sets of points in Fig. 6.19 reflect the progressive loss of predictive accuracy in the NWP model at longer lead times. As the scatter of points increases at longer projections, the slopes of the MOS forecast equations become more horizontal, leading to temperature forecasts that are more like the climatological temperature, on average. This is reasonable and desirable, since as the NWP model provides less information about the future state of the atmosphere, temperature forecasts differing substantially from the climatological average temperature are less justified. In the limit of arbitrarily long forecast projection, an NWP model will really provide no more information than will the climatological value of the predictand, the slope of the corresponding MOS equation would be zero, and the appropriate temperature forecast consistent with this (lack of) information would simply be the climatological average temperature. Thus, it is sometimes said that "MOS converges toward the climatology." By contrast, the perfect-prog equation (dashed lines, reproduced from Fig. 6.18) take no account of the decreasing accuracy of the NWP model at longer projections, and continue to produce temperature forecasts as if the thickness forecasts

Table 6.7

Example MOS Forecasts Produced by U.S. National Meteorological Center for Albany, New York, Shortly After 0000 UTC on March 12, 1993[a]

FOUS14	KWBC	120000																	
ALB	EC	NGM	MOS	GUIDANCE			3/12/93		0000		UTC								
DAY		/MAR	12				/MAR	13							/MAR	14			
HOUR		06	09	12	15	18	21	00	03	06	09	12	15	18	21	00	03	06	09	12
MX / MN								32				14				33				23
TEMP		20	16	14	22	30	31	23	18	17	17	19	28	31	32	30	26	24	25	25
DEWPT		10	7	4	3	3	4	7	7	9	10	12	18	21	23	26	26	24	25	25
CLDS		BK	SC	CL	CL	SC	SC	CL	SC	SC	OV	OV	OV	OV	OV	OV	OV	OV	OV	OV
WDIR		30	30	31	31	30	30	31	32	00	12	08	06	04	03	36	34	34	33	32
WSPD		08	08	07	09	09	07	03	01	00	03	03	05	10	18	13	23	32	29	25
POP06				0		0		0		7		42		72		98		100		92
POP12								0				37				100				100
QPF				0/		0/		0/0		0/		1/1		4/		5/6		2/		2/3
TSV06			0	/ 3	0	/ 3	0	/ 3	1	/ 4	0	/ 6	4	/ 0	14	/ 0	21	/ 1	23	/ 0
TSV12					0	/ 4			1	/ 5			3	/ 5			26	/ 0		
PTYPE		S	S	S	S	S	S	S	S	S	S	S		S		S		S		S
POZP		0	0	3	1	0	1	0	2	0	1	0		3		3		4		0
POSN		94	100	97	99	100	99	100	98	100	99	98		96		90		86		95
SNOW				0/		0/		0/0		0/		1/1		2/		4/6		2/		2/4
CIG		7	7	7	7	7	7	7	7	7	6	5		4		3				
VIS		5	5	5	5	5	5	5	5	4	4	2		1		1				
OBVIS		N	N	N	N	N	N	N	N	N	N	N		F		F				

[a] Various weather elements are forecast, at projections up to 60 h and at intervals as close as 3 h. Detailed explanations of all the forecast elements are given in Dallavalle *et al.* (1992).

were perfect. Figure 6.19 emphasizes that the result is overconfident temperature forecasts, with both very warm and very cold temperatures forecast much too frequently.

6.4.4 Operational MOS Forecasts

Interpretation and extension of dynamical NWP forecasts through MOS systems have been implemented at a number of national meteorological centers, including those in the Netherlands (Lemcke and Kruizinga, 1988), Britain (Francis *et al.*, 1982), Italy (Conte *et al.*, 1980), China (Lu, 1991), Spain (Azcarraga and Ballester, 1991), Canada (Brunet *et al.*, 1988), and the United States (Carter *et al.*, 1989).

MOS forecast products can be quite extensive, as illustrated by Table 6.7, which shows the "FOUS14 panel" (Dallavalle *et al.*, 1992) of March 12, 1993 for Albany, New York. Forecasts for a wide variety of weather elements are provided, at projections up to 60 h and at intervals as close as 3 h. After lines indicating the dates and times [Universal Time Coordinated (UTC)] are forecasts for daily maximum and minimum temperatures; temperatures, dewpoint temperatures, cloud coverage, windspeed, and wind direction at 3-h intervals; probabilities of measurable precipitation at 6- and 12-h intervals; forecasts for precipitation amount; thunderstorm probabilities; precipitation type; freezing and frozen precipitation probabilities; snow amount; and ceiling, visibility, and obstructions to visibility. FOUS14 forecasts are issued twice daily for each of several hundred locations in the United States.

The MOS equations underlying the FOUS14 forecasts are *seasonally stratified*, usually with separate forecast equations for the "warm season" (April through September) and "cool season" (October through March). This two-season stratification allows the MOS forecasts to incorporate different relationships between predictors and predictands at different times of the year. A finer stratification (3-month seasons, or separate month-by-month equations) would probably be preferable if sufficient developmental data were available.

The forecast equations for all elements except temperatures, dewpoints, and winds are *regionalized*; that is, developmental data from groups of nearby and climatically similar stations were composited in order to increase the sample size when deriving the forecast equations. For each regional group, then, forecasts are made with the same equations and the same regression coefficients. This does not mean that the forecasts for all the stations in the group are the same, however, since interpolation of the NWP output to the different forecast locations yields different values for the predictors from the NWP model. Some of the MOS equations also contain predictors representing local climatological values, which introduce further differences in the forecasts for the nearby stations. Regionalization is especially valuable for producing good forecasts of rare events.

In order to enhance consistency among the forecasts, some of the MOS equa-

tions are developed simultaneously for several of the forecast elements. This means that the same predictor variables, although with different regression coefficients, are forced into prediction equations for related predictands in order to enhance the consistency of the forecasts. For example, it would be physically unreasonable and clearly undesirable for the forecast dewpoint to be higher than the forecast temperature. To help ensure that such inconsistencies appear in the forecasts as rarely as possible, the MOS equations for maximum temperature, minimum temperature, and the 3-h temperatures and dewpoints all contain the same predictor variables. Similarly, the four groups of forecast equations for windspeeds and directions, the 6- and 12-h precipitation probabilities, the 6- and 12-h thunderstorm probabilities, and the probabilities for precipitation type were also developed simultaneously to enhance their consistency.

6.5 Probabilistic Field (Ensemble) Forecasts

In Section 1.3 it was asserted that "dynamical chaos" ensures that the future behavior of the atmosphere can not be known with certainty. Lorenz (1993) has provided a very readable exposition of this problem. Because the atmosphere can never be completely observed, either in terms of spatial coverage or accuracy of measurements, a fluid-dynamical numerical weather prediction (NWP) model of its behavior will always begin calculating forecasts from a state at least slightly different from that of the real atmosphere. These models (and other nonlinear dynamical systems, including the real atmosphere) exhibit the property that solutions (forecasts) started from only slightly different initial conditions will yield quite different results for projections sufficiently far into the future. For synoptic-scale predictions using NWP models, "sufficiently far" is a matter of days or (at most) weeks, so that this problem of sensitivity to initial conditions is of practical importance for weather forecasting.

NWP models are the mainstay of weather forecasting, and the inherent uncertainty of their results must be appreciated and quantified if their information is to be utilized most effectively. For example, a single categorical forecast of the hemispheric 500-mb-height field 2 days in the future is only one member of an essentially infinite collection of 500-mb-height fields that could occur. Even if this categorical forecast is the best possible single forecast that can be constructed, its usefulness and value will be enhanced if aspects of the probability distribution of which it is a member can be estimated and communicated. This is the problem of probabilistic field forecasting.

Probability forecasts of a scalar quantity, such as a maximum daily temperature at a single location, are relatively straightforward. Many aspects of producing such forecasts have been discussed in this chapter, and the uncertainty of such forecasts can be expressed using univariate probability distributions of the kind described in Chapter 4. However, probabilistic field forecasting is a much bigger and more difficult problem. A single atmospheric field, for example that of hemispheric

500-mb height, might be represented by the values of thousands of 500-mb heights at regularly spaced locations, or gridpoints (e.g., Hoke *et al.*, 1989). Construction of forecasts including probabilities for all of these heights and their relationships (e.g., correlations) with heights at the other gridpoints is a very big task, and in practice only approximations to a complete probability description have been achieved. A further difficulty is expressing and communicating aspects of the large amounts of information in a probabilistic field forecast.

6.5.1 Stochastic Dynamic Forecasts

Much of the conceptual basis for probabilistic field forecasts is drawn from Epstein (1969c), who presented both theoretical and practical approaches to the problem of uncertainty in NWP forecasts. In this approach, which Epstein called *stochastic dynamic prediction*, the physical laws governing the motions and evolution of the atmosphere are regarded as deterministic. However, in practical problems the equations that describe these laws must operate on initial values that are not known with certainty, and that can therefore be described by a probability distribution. Conventional deterministic forecasts use the governing equations to describe the future evolution of a single initial state that is assumed to be the true initial state. The idea behind stochastic dynamic forecasts is to allow the deterministic governing equations to operate on the probability distribution describing the uncertainty about the initial state of the atmosphere. This process yields, as forecasts, probability distributions describing uncertainty about the future state of the atmosphere. (Actually, since NWP models are not perfect representations of the real atmosphere, these imperfections further contribute to forecast uncertainty.)

Visualizing the initial and forecast probability distributions is difficult, especially when they involve joint probabilities pertaining to large numbers of variables. This visualization is most commonly and easily done using the phase-space concept. A *phase space* is a geometric representation of a dynamical system, where each of the coordinate axes defining this geometry pertains to one of the variables in the system.

For example, a simple dynamical system that is commonly encountered in textbooks on physics or differential equations is the swinging pendulum. The state of a pendulum can be completely described by two variables: its angular position and its velocity. At the extremes of the pendulum's arc, its angular position is maximum (positive or negative) and its speed is zero. At the bottom of its arc the angular position of the swinging pendulum is zero and its velocity is maximum. When the pendulum finally stops, both its angular position and velocity are zero. Because the motions of a pendulum can be described by two variables, its phase space is two-dimensional. That is, its phase space is a *phase-plane*. The changes through time of the state of the pendulum system can be described by a path, known as an *orbit*, or *trajectory*, on this phase-plane.

Figure 6.20 shows the trajectory of a hypothetical pendulum in its phase space. That is, this figure is a graph in phase space of the motions of a pendulum, and

Fig. 6.20 Trajectory of a swinging pendulum in its two-dimensional phase space, or phase-plane. The pendulum has been dropped from a position on the right, from which point it swings in arcs of decreasing angle. Finally, it slows to a stop with zero velocity in the vertical position.

their changes through time. The trajectory begins at the single point corresponding to the initial state of the pendulum: It is dropped from the right with initially zero velocity. As it drops it accelerates and acquires leftward velocity, which increases until the pendulum passes through the vertical position. The pendulum then decelerates, slowing until it stops at its maximum left position. As the pendulum drops again it moves to the right, stopping short of its initial position because of friction. The pendulum continues to swing back and forth until it finally comes to rest in the vertical position.

The phase space of an atmospheric model has many more dimensions than that of the pendulum system. Epstein (1969c) considered a highly simplified model of the atmosphere having only eight variables. Its phase space was therefore eight-dimensional, which is small but still much too big to imagine explicitly. The phase spaces of operational NWP models typically have millions of dimensions, each corresponding to one of the millions of variables represented in these models. The trajectory of the atmosphere or a model of the atmosphere is also qualitatively more complicated than that of the pendulum because it is not "attracted" to a single point in the phase space, as is the trajectory of the pendulum. Nevertheless, the changes in the flow in a model atmosphere through time can still be imagined abstractly as a trajectory through this multidimensional space.

The uncertainty about the initial state of the atmosphere, from which a NWP

model is initialized, can be conceived of as a probability distribution in its phase space. In a two-dimensional phase space, one might imagine a bivariate normal distribution (Chapter 4), with ellipses of constant probability describing the spread of plausible states of the atmosphere around the "best guess," or mean value. Alternatively, one can imagine a cloud of points around the mean value, whose density decreases with distance from the mean. In a three-dimensional phase space the distribution might be imagined as a cigar- or egg-shaped cloud of points, again with density decreasing with distance from the mean value. Higher-dimensional spaces cannot be visualized explicitly, but probability distributions within them can be imagined by analogy.

In concept, a stochastic dynamic forecast moves the probability distribution of the initial state through the phase space as the forecast is advanced in time, according to the laws of fluid dynamics represented in the NWP model equations. However, trajectories in the phase space of a NWP model (or of the real atmosphere) are not nearly as smooth and regular as the pendulum trajectory shown in Fig. 6.20. As a consequence, the shape of the initial distribution is stretched and distorted as the forecast is advanced. It will also become more dispersed at longer forecast projections, reflecting the increased uncertainty of forecasts further into the future.

Equations describing this evolution of the probability distribution can be derived through introduction of a continuity, or conservation equation for probability (Ehrendorfer, 1994). However, the dimensions of phase spaces for problems of practical forecasting interest are too large to allow direct solution of these equations. Epstein (1969c) introduced a simplification that rests on a restrictive assumption about the shapes of probability distributions in phase space, which is expressed in terms of the moments of the distribution. However, even this approach is impractical for all but the simplest atmospheric models.

6.5.2 Ensemble Forecasting

The practical solution to the analytic intractability of sufficiently detained stochastic dynamic equations is to approximate these equations using Monte Carlo methods. These Monte Carlo solutions bear the same relationship to stochastic dynamic forecast equations as the Monte Carlo resampling tests described in Section 5.3 bear to the analytical tests they approximate. Recall that resampling tests are appropriate and useful in situations where the underlying mathematics are difficult or impossible to evaluate analytically.

The ensemble forecast procedure begins by drawing a finite sample from the probability distribution describing the uncertainty of the initial state of the atmosphere. One can imagine that a few members of the point cloud surrounding the mean atmospheric state in phase space are picked randomly. These points are called the *ensemble* of initial conditions, and each represents a plausible initial state of the atmosphere consistent with the uncertainties in observation and analy-

sis. Rather than explicitly predict the movement of the initial-state probability distribution through phase space, that movement is approximated by the collective trajectories of the ensemble of sampled initial points. It is for this reason that the Monte Carlo approximation to stochastic dynamic forecasting is known as *ensemble forecasting*. Each point in the initial ensemble provides the initial conditions for a separate run of the NWP model. At this initial time, all of the ensemble members are very similar to each other. The distribution in phase space of this ensemble of points after the forecasts have been advanced to a future time then approximates how the full true initial probability distribution would have been transformed by the governing physical laws. Leith (1974) outlines the procedure nicely.

Figure 6.21 illustrates the nature of ensemble forecasting in an idealized two-dimensional phase space. The ⊗ in the initial-time ellipse represents the single "best" initial value, from which a conventional deterministic NWP forecast would begin. The evolution of this single forecast in the phase space, through an intermediate forecast projection and to a final forecast projection, is represented by the heavy solid lines. However, the position of this point in phase space at the

Fig. 6.21 Schematic illustration of concepts in ensemble forecasting, plotted in terms of a two-dimensional phase space. The heavy line represents the evolution of the single "best" analysis of the initial state of the atmosphere, corresponding to the more traditional single deterministic forecast. The dashed lines represent the evolution of other individual ensemble members. The ellipse in which they originate represents the statistical distribution of initial atmospheric states, which are very close to each other. At the intermediate projection, all of the ensemble members are still reasonably similar. By the time of the final projection some of the ensemble members have undergone a regime change, and represent qualitatively different flows. Any of the ensemble members, including the solid line, are plausible trajectories for the evolution of the real atmosphere, and there is no way of knowing in advance which will represent the real atmosphere most closely.

initial time represents only one of the many plausible initial states of the atmosphere consistent with errors in the analysis. Clustering around it are other plausible states, which represent the probability distribution of states of the atmosphere at the initial time. This distribution is represented by the small ellipse. The open circles in this ellipse represent eight other members of this distribution. This ensemble of nine initial states approximates the variations represented by the full distribution from which they were drawn.

The Monte Carlo approximation to the stochastic dynamic forecast is constructed by repeatedly running the NWP model, once for each member of the initial ensemble. The trajectories through the phase space of each ensemble member are only modestly different at first, indicating that all nine NWP integrations are producing fairly similar forecasts at the intermediate projection. Accordingly, the probability distribution describing uncertainty about the state of the atmosphere at the intermediate projection is not much larger than at the initial time. However, between the intermediate and final projections the trajectories diverge markedly, with three (including the one started from the mean value of the initial distribution) producing forecasts that are similar to each other, and the remaining six members of the ensemble predicting a rather different atmospheric state at that time. The underlying distribution of uncertainty that was fairly small at the initial time has been stretched substantially, as represented by the large ellipse at the time of the final projection. The dispersion of the ensemble members at that time allows the nature of that distribution to be estimated, and is indicative of the uncertainty of the forecast. If only the single forecast started from the "best" initial condition had been made, this information would not be available.

6.5.3 Choosing Initial Ensemble Members

Ideally, one would like to produce ensemble forecasts based on a large number of possible initial atmospheric states. However, each member of an ensemble of forecasts is produced by a complete rerunning of the NWP model and requires a substantial amount of computing. As a practical matter, computer time is a limiting factor at operational forecast centers, and the sizes of operational forecast ensembles are also limited. Consequently, it is important that initial ensemble members be chosen well.

The simplest, and historically first, method of generating initial ensemble members is to begin with a "best" analysis asssumed to be the mean of the probability distribution representing the uncertainty of the initial state of the atmosphere. Variations around this mean state can be easily generated, by adding random numbers characteristic of the errors or uncertainty in the instrumental observations underlying the analysis (Leith, 1974). For example, these random values might be Gaussian variates with zero mean, implying unbiased measurement errors.

In practice, however, simply adding random numbers to a single initial field has been found to yield ensembles whose members are too similar to each other. That

is, the variability of the resulting forecast ensemble appears to underestimate the uncertainty in the forecast. Mureau *et al.* (1993) and Toth and Kalnay (1993) describe more elaborate alternatives, which were in operational use as of 1993. These methods attempt to include initial states in the ensemble whose trajectories are expected to diverge more strongly in phase space.

Another approach to the choice of initial states is known as *lagged-average forecasting* (Hoffman and Kalnay, 1983). This alternative is attractive operationally because it requires no additional computing time beyond what is necessary for the routine forecasting cycle. The idea here is to take the ensemble of initial conditions from previously run forecasts, which are valid at the initial time of the new ensemble forecast. That is, if a new ensemble forecast is to be started on Friday, initial members of the ensemble would include Thursday's forecast for Friday, Wednesday's forecast for Friday, Tuesday's forecast for Friday, and so on. However, these ensemble members require no additional computing time, because they have already been run on the days for which they were initialized. If, each day, the NWP model is projected ten days into the future, there will be four forecasts pertaining to day 7. These will consist of the 7-day projection from the current run, through the 10-day projection of the forecast initialized 3 days previously.

6.5.4 Ensemble Averaging

One application of ensemble forecasting is to average the members of the ensemble to obtain a single forecast. The motivation is to obtain a forecast that is more accurate than the single forecast initialized with the best estimate of the initial state of the atmosphere. Epstein (1969a) pointed out that the time-dependent behavior of the ensemble mean is different from the solution of forecast equations using the initial mean value, and concluded that in general the best forecast is not the single forecast initialized with the best estimate of initial conditions. This conclusion should not be surprising, since a NWP model is in effect a highly nonlinear function that transforms a set of initial atmospheric conditions to a set of forecast atmospheric conditions.

In general, the average of a nonlinear function over some set of particular values of its argument is not the same as the function evaluated at the average of those values. That is, if the function $f(x)$ is nonlinear,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \neq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right). \tag{6.29}$$

To illustrate simply, consider the three values $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. For the nonlinear function $f(x) = x^2 + 1$, the left side of Eq. (6.29) is $5\frac{2}{3}$, and the right side of that equation is 5. One can easily verify that the inequality of Eq. (6.29) holds for other nonlinear functions [e.g., $f(x) = \log(x)$ or $f(x) = 1/x$] as well. [By contrast, for the linear function $f(x) = 2x + 1$ the two sides of Eq. (6.29) are both equal to 5.]

Extending this idea to ensemble forecasting, one might like to know the atmospheric state corresponding to the center of the ensemble in phase space for some time in the future. This central value of the ensemble will approximate the center of the stochastic-dynamic probability distribution at that future time, after the initial distribution has been transformed by the forecast equations. The Monte Carlo approximation to this future value is the ensemble average forecast. The ensemble average forecast is obtained simply by averaging together the ensemble members for the time being forecast, which corresponds to the left side of Eq. (6.29). By contrast, the right side of Eq. (6.29) represents the single forecast started from the average initial value of the ensemble members. Depending on the nature of the initial distribution and the dynamics of the NWP model, this single forecast may or may not be close to the ensemble average forecast.

In the context of weather forecasts, the benefits of ensemble averaging appear to derive primarily from averaging out variations in the short (synoptic) waves between members of the ensemble, while retaining the more slowly varying long-wave pattern (Leith, 1974). Operationally, however, ensemble averaging has not provided much increase in skill for forecasts beyond the short range (J. M. Murphy, 1990; Toth and Kalnay, 1993). Palmer (1993) suggests that ensemble averaging will improve the forecast only until a *regime change*, or a change in the long-wave pattern, and he illustrates this concept nicely using the simple Lorenz (1963) model. This problem is also illustrated in Fig. 6.21, where a regime change is represented by the bifurcation of the trajectories of the ensemble members between the intermediate and final forecast projections. At the intermediate projection, before some of the ensemble members undergo this regime change, the center of the distribution of ensemble members is well represented by the ensemble average, which is a better "central" value than is the single member of the ensemble started from the "best" initial condition. At the final forecast projection the distribution of states has been distorted into two distinct groups. Here the ensemble average will be located somewhere in the middle, but near none of the ensemble members.

6.5.5 Dispersion of the Ensemble

A particularly valuable aspect of ensemble forecasting is its capacity to yield information about the magnitude and nature of the uncertainty in a forecast. Qualitatively, one has more confidence that the ensemble mean is close to the eventual state of the atmosphere if the dispersion of the ensemble is small. Conversely, if the ensemble members are all very different from one another, the future state of the atmosphere is very uncertain. One approach to "forecasting forecast skill" is to anticipate the skill of a forecast as being inversely related to the dispersion of ensemble members. Operationally, forecasters do this informally when comparing the results from different NWP models, or when comparing successive forecasts for a particular time in the future that were initialized on different days.

Because of the large amount of information in an ensemble of atmospheric

fields, quantitative portrayal of the uncertainty implied by an ensemble forecast, in a way that is informative to the forecaster, poses a difficult problem. The most promising means of expressing the uncertainty implied by the dispersion of the ensemble members is through the use of graphical summaries. As of this writing, there has been little enough operational experience with ensemble forecasting that no particular set of summaries is standard. However, some interesting suggestions have been offered in the literature.

Gleeson (1967) suggested combining maps of forecast u and v wind components with maps of probabilities that the forecasts will be within 10 knots of the eventual observed values. Epstein and Fleming (1971) suggested that a probabilistic depiction of a horizontal wind field could take the form of Fig. 6.22. Here the lengths and orientations of the lines indicate the ensemble mean forecast wind vectors, blowing from the gridpoints toward the directions of the ellipses. The probability is 0.50 that the true wind vectors will terminate within the corresponding ellipse. It has been assumed in this figure that the uncertainty in the wind forecasts is described by the bivariate normal distribution, and the ellipses have been drawn as discussed in Example 9.4. The tendency for the ellipses to be oriented in a north–south direction indicates that the uncertainties of the meridional winds are greater than those for the zonal winds, and the tendency for the larger velocities to be associated with larger ellipses indicates that these wind values are more uncertain.

Epstein and Fleming (1971) also suggested use of side-by-side plots of an ensemble average field, and the corresponding field of standard deviations. Portions of the standard deviation field that are small indicate that the corresponding re-

Fig. 6.22 A forecast wind field expressed in probabilistic form. Line lengths and orientations show forecast mean wind vectors directed from the gridpoint locations to the ellipses. The ellipses indicate boundaries containing the observed wind with probability 0.50. [From Epstein and Fleming (1971), reproduced with permission of the American Meteorological Society.]

gions of the forecast mean field are relatively certain, while large standard deviations indicate regions where relatively little confidence can be placed in the mean forecast. Figure 6.23 illustrates this format for an operational ensemble forecast of the Northern Hemisphere 500-mb-height field. Panel a shows the ensemble mean forecast field, and panel b shows the sample standard deviation of the forecast heights over 14 ensemble members, scaled by the climatological standard deviation at each gridpoint. The low standard deviations over Greenland and northern Russia indicate good agreement among the ensemble members about the placement and intensities of the closed lows located there, while the rather high

Fig. 6.23 Ensemble average forecast 500-mb-height field for the Northern Hemisphere (a) and the field of standard deviations of heights among the ensemble members, scaled by the standard deviations of the climatological 500-mb heights (b). Low standard deviations indicate regions of relative agreement among the ensemble members, while flow features in regions of high standard deviation are more uncertain. [From Tracton and Kalnay (1993), reproduced with permission of the American Meteorological Society.]

b

Fig. 6.23 *Continued.*

standard deviation in the northwest Pacific indicates that the forecast trough there is quite uncertain.

Other kinds of probabilistic field maps are also possible, many of which may be suggested by the needs of particular forecast applications. Figure 6.24, showing relative frequencies of forecast 1000–500-mb thickness over North America being less than 5400 m, is one such possibility. Forecasters often use this thickness value as the dividing line between rain and snow. At each gridpoint, the forecast thicknesses from 14 ensemble members have been ranked to form an empirical cumulative probability distribution, and the probability corresponding to 5400 m thickness at each of the points has been extracted and plotted. Clearly, similar maps for other thickness values could be constructed as easily. Figure 6.24 indi-

Fig. 6.24 Forecast probabilities that the 1000–500-mb thicknesses will be less than 5400 m over North America, as estimated from the relative frequencies of thicknesses in fourteen ensemble forecast members. [From Tracton and Kalnay (1993), reproduced with permission of the American Meteorological Society.]

cates a relatively high confidence that the cold-air outbreak in the eastern United States will bring air sufficiently cold to produce snow as far south as the Gulf Coast.

6.6 Subjective Probability Forecasts

Most of this chapter has dealt with the derivation of *objective* forecasts, or forecasts produced by means that are automatic. Objective forecasts are determined unambiguously by the nature of the forecasting procedure, and the values of the variables that are used to drive it. However, objective forecasting procedures necessarily rest on a number of subjective judgments made during their development. Nevertheless, some people feel more secure with the results of objective forecasting procedures, apparently taking comfort from their lack of contamination by the vagaries of human judgment. Apparently, such individuals feel that objective forecasts are in some way less uncertain than human-mediated forecasts.

One very important—and perhaps irreplaceable—role of human forecasters in

the forecasting process is in the subjective integration and interpretation of objective forecast information. These objective forecast products are often called forecast *guidance*, and include deterministic forecast information from NWP integrations and statistical guidance from MOS systems or other interpretive statistical products. Human forecasters also use, and incorporate into their judgments, available atmospheric observations (surface maps, radar charts, etc.), and prior information ranging from persistence or simple climatological statistics, to their individual previous experiences with similar meteorological situations. The result is (or should be) a forecast reflecting, to the maximum practical extent, a forecaster's state of knowledge about the future evolution of the atmosphere.

Human forecasters can rarely, if ever, fully describe or quantify their personal forecasting processes. Thus, the distillation by a human forecaster of disparate and sometimes conflicting information is known as *subjective* forecasting. A subjective forecast is one formulated on the basis of the judgment of one or more individuals. Making a subjective weather forecast is a challenging process precisely because future states of the atmosphere are inherently uncertain. The uncertainty will be larger or smaller in different circumstances—some forecasting situations are more difficult than others—but it will never really be absent.

Since the future states of the atmosphere are inherently uncertain, a key element of good and complete subjective weather forecasting is the reporting of some measure of the forecaster's uncertainty. It is the forecaster who is most familiar with the atmospheric situation, and it is therefore the forecaster who is in the best position to evaluate the uncertainty associated with a given forecasting situation. While it is common to issue categorical forecasts (i.e., forecasts containing no expression of uncertainty) such as "tomorrow's maximum temperature will be 27° F," an individual issuing this forecast would not seriously expect the temperature to be exactly 27° F. Given a forecast of 27° F, temperatures of 26 or 28° F would generally be regarded as nearly as likely, and in this situation the forecaster would usually not really be surprised to see tomorrow's maximum temperature anywhere between 25 and 30° F.

While uncertainty about future weather can be reported verbally, using phrases such as "chance" or "likely," such qualitative descriptions are open to different interpretations by different people (e.g., Murphy and Brown, 1983). Even worse, however, is the fact that such qualitative descriptions do not precisely reflect the forecaster's uncertainty about, or degree of belief in, the future weather. The forecaster's state of knowledge is most accurately reported, and the needs of the forecast user are best served, if the intrinsic uncertainty is quantified in probability terms. Thus, the Bayesian view of probability as the degree of belief of an individual holds a central place in subjective forecasting. Note that since different forecasters have somewhat different information on which to base their judgments (e.g., different sets of experiences with similar past forecasting situations), it is perfectly reasonable to expect that their probability judgments may differ somewhat as well.

6.6.1 The Subjective Distribution

Before reporting a subjective degree of uncertainty as part of a forecast, a forecaster needs to have an image of that uncertainty. The information about one's uncertainty can be thought of as residing in one's subjective distribution for the event in question. The *subjective distribution* is a probability distribution in the same sense as the theoretical distributions described in Chapter 4. Sometimes, in fact, one of the theoretical distributions in Chapter 4 may provide a very good approximation to one's subjective distribution. Subjective distributions are interpreted from a Bayesian perspective as the quantification of an individual's degree of belief in each possible outcome for the event to be forecast.

While preparing to make a forecast, a forecaster internally develops a subjective distribution. The possible weather outcomes are subjectively weighed, and an internal judgment is formed as to their relative likelihoods. This process occurs whether or not the forecast is to be a probability forecast, or indeed whether the forecaster is even aware of the process. However, unless one believes that uncertainty can somehow be expunged from the process of weather forecasting, it should be clear that better forecasts will result when forecasters think explicitly about their subjective distributions and the uncertainty that those distributions describe.

It is easiest to approach the concept of subjective probabilities with a familiar but simple example. Subjective probability-of-precipitation (PoP) forecasts have been routinely issued in the United States since 1965. These forecasts specify the probability that measurable precipitation (i.e., at least 0.01 in.) will occur at a particular location during a specified time period. The forecaster's subjective distribution for this event is so simple that one might not notice that it is a probability distribution. However, the events "precipitation" and "no precipitation" divide the sample space into two MECE events. The distribution of probabilities over these events is discrete, and consists of two elements: one probability for the event "precipitation" and another probability for the event "no precipitation." This distribution will be different for different forecasting situations, and perhaps for different forecasters assessing the same situation. However, the only thing about a forecaster's subjective distribution for the PoP that can change from one forecasting occasion to another is the probability, and this will be different to the extent that the forecaster's degree of belief regarding future precipitation occurrence is different. The PoP ultimately issued by the forecaster should be the forecaster's subjective probability for the event "precipitation," or perhaps a suitably rounded version of that probability. That is, the forecaster's job is to evaluate the uncertainty associated with the possibility of future precipitation occurrence, and to report that uncertainty to the users of the forecasts.

An operational subjective probability forecasting format that is only slightly more complicated than the format for PoP forecasts is that for the 30- and 90-day outlooks for temperature and precipitation issued by the Climate Analysis Center

of the U.S. National Weather Service (Epstein, 1988; Wagner, 1989). Average temperature and total precipitation at a given location over the upcoming 30- or 90-day period are continuous variables, and complete specifications of the forecaster's uncertainty regarding their values would require continuous subjective probability distribution functions. The format for these forecasts is simplified considerably by defining three-event MECE partitions of the sample spaces for the temperature and precipitation outcomes, thereby transforming the problem into one of specifying a discrete probability distribution with three possible outcomes.

The three events are defined with respect to the local climatological distributions for the 30- or 90-day period to which the forecast pertains, in a manner similar to that used in reporting recent climate anomalies for these two variables (Examples 4.5 and 4.7). In the case of temperature, a "cold" outcome is one that would fall in the lowest 30% of the climatological temperature distribution, a "near-normal" outcome is one that would fall in the middle 40% of that climatological distribution, and a "warm" outcome would fall in the upper 30% of the climatological distribution. Thus, the continuous temperature scale is divided into three discrete events according to the 30th and 70th percentiles of the local climatological temperature distribution. Similarly, the precipitation outcomes are defined with respect to the same quantiles of the local precipitation distribution, so "dry," "near-normal," and "wet" precipitation outcomes correspond to the driest 30%, middle 40%, and wettest 30% of the climatological distribution, respectively.

It is often not realized that these 30- and 90-day forecasts are probability forecasts. That is, the forecast quantities are probabilities for the events defined in the previous paragraph rather than temperatures or precipitation amounts per se. Through experience the forecasters have found that future periods with other than the climatological relative frequency of the near-normal events cannot be discerned. That is, regardless of the probability that forecasters have assigned to the near-normal event, the subsequent relative frequencies are nearly 40%. Therefore, operationally the probability of 0.4 is always assigned to the middle category. Since the probabilities for the three MECE events must all add to one, this restriction on the near-normal outcomes implies that each full forecast distribution can be specified by a single probability. In effect, the forecaster's subjective distribution for the upcoming month or season is collapsed to a single probability. Forecasting a 10% chance that next month will be cool implies a 50% chance that next month will be warm, since the format of the forecasts always carries the implication that the chance of a near-normal event is 40%.

One benefit of this restriction is that the full 30- or 90-day forecast of the temperature or precipitation outcome for the entire country can be displayed on a single map. Figure 6.25 shows the forecasts made for January 1987. The mapped quantities are probabilities in percentage terms for the more likely of the two extreme events in each case ("cool" or "warm," and "dry" or "wet," respectively). The smallest contours—30%—thus indicate locations where the climatological probabilities are forecast, since by construction each local climatological distri-

Fig. 6.25 Thirty-day outlooks for temperature and precipitation over North America for January 1987, issued by the U.S. National Weather Service. Contoured quantities are probabilities in percentage form for the more likely of the two extreme categories—"cool" or "warm," and "dry" or "wet." The 30% contour indicates locations where the three climatological probabilities (0.3, 0.4, 0.3) are forecast.

bution yields probabilities for the lower, middle, and upper categories of 0.3, 0.4, and 0.3, respectively. Regions for which the probabilities are larger for one tail or the other are indicated on the maps by the labels "warm" or "cold," and "light" or "heavy," respectively. Since precipitation is more difficult to forecast, less extreme forecast deviations from the climatological distribution are generally issued for this variable.

6.6.2 Credible Interval Forecasts

It has been argued here that inclusion of some measure of the forecaster's uncertainty is desirable in any weather forecast. Historically, resistance to this idea has been based in part on the practical consideration that the forecast format should be compact and easily understandable. In the case of PoP forecasts the subjective distribution is sufficiently simple that it can be reported with a single number, and is no more cumbersome than issuing categorical forecasts of "precipitation" or "no precipitation." When the subjective distribution is continuous, however, some approach to sketching its main features is a practical necessity if its probability information is to be conveyed in a publicly issued forecast. "Discretizing" the subjective distribution, as is done for the 30- and 90-day outlooks, is one approach to simplifying a continuous subjective probability distribution in terms of one or a few easily expressible quantities. Alternatively, if the forecaster's subjective distribution on a given occasion can be reasonably well approximated by one of the theoretical distributions described in Chapter 4, another approach to simplifying

its communication would be to report the parameters of the approximating distribution. There is no guarantee, however, that subjective distributions will always (or even ever) correspond to a familiar theoretical form.

One very attractive and workable alternative for introducing probability information into the forecasts for continuous meteorological variables is the use of *credible-interval* forecasts. This forecast format is used operationally in Sweden (Ivarsson *et al.*, 1986), but to date has been used only experimentally in the United States (Murphy and Winkler, 1974; Peterson *et al.*, 1972; Winkler and Murphy, 1979). In unrestricted form, a credible-interval forecast requires specification of three quantities: two points defining an interval of the forecast variable, and a probability (according to the forecaster's subjective distribution) that the forecast quantity will fall in the designated interval. Often the requirement is made that the credible interval be located in the middle of the subjective distribution. In this case the specified probability is equally distributed on either side of the median, and the forecast is called a *central credible-interval* forecast.

There are two special cases of the credible-interval forecast format, each requiring that only two quantities be forecast. The first is the *fixed-width* credible-interval forecast. As the term implies, the width of the credible interval is the same for all forecasting situations and is specified in advance for each predictand. Thus the forecast includes a location for the interval, generally specified as its midpoint, and a probability that the outcome will occur in the forecast interval. For example, the Swedish credible-interval forecasts for temperature are of the fixed-width type, with the interval size specified to be $\pm 3°C$ around the midpoint temperature. These forecasts thus include a forecast temperature, together with a probability that the subsequently observed temperature will be within 3° C of the forecast temperature. The two forecasts "15° C, 90%" and "15° C, 60%" would both indicate that the forecaster expects the temperature to be about 15° C, but the inclusion of probabilities in the forecasts shows that much more confidence can be placed in the former as opposed to the latter of the two forecasts of 15° C. Because the forecast interval is "central," these two forecasts would also imply 5 and 20% chances, respectively, for the temperature to be colder than 12° and warmer than 18°.

Some forecast users would find the unfamiliar juxtaposition of a temperature and a probability in a fixed-width credible-interval forecast to be somewhat jarring. An alternative forecast format that could be implemented more subtly is the *fixed-probability* credible-interval forecast. In this format, it is the probability contained in the forecast interval, rather than the width of the interval, that is specified in advance and is constant from forecast to forecast. This makes the probability part of the credible-interval forecast implicit, so the forecast consists of two numbers having the same physical dimensions as the quantity being forecast.

Figure 6.26 illustrates the relationship of 75% central credible intervals for two subjective distributions having the same location. The shorter, broader distribution represents a relatively uncertain forecasting situation, where events fairly far

Fig. 6.26 Two hypothetical subjective distributions shown as probability density functions (PDFs). The two distributions have the same location, but reflect different degrees of uncertainty. The tall, narrow distribution represents an easier (less uncertain) forecasting situation, and the broader distribution represents a more difficult forecast problem. Arrows delineate 75% central credible intervals in each case.

from the center of the distribution are regarded as having substantial probability. A relatively wide interval is therefore required to subsume 75% of this distribution's probability. On the other hand, the tall and narrow distribution reflects considerably less uncertainty, and a much narrower forecast interval contains 75% of its density. If the variable being forecast is temperature, the 75% credible-interval forecasts for these two cases might be "10° to 20°," and "14° to 16°," respectively.

A strong case can be made for operational credible-interval forecasts (Murphy and Winkler, 1974, 1979). Since nonprobabilistic temperature forecasts are already often specified as ranges, fixed-probability credible-interval forecasts could be introduced into forecasting operations quite unobtrusively. Forecast users not wishing to take advantage of the implicit probability information would notice little difference from the present forecast format, while those understanding the meaning of the forecast ranges would derive additional benefit. Even forecast users unaware that the forecast range is meant to define a particular interval of fixed probability might notice over time that the interval widths were related to the precision of the forecasts.

6.6.3 Assessing Discrete Probabilities

Experienced weather forecasters are able to formulate subjective probability forecasts that apparently quantify their uncertainty regarding future weather quite successfully. Examination of the error characteristics of such forecasts (Chapter 7) reveals that they are largely free of the biases and inconsistencies sometimes ex-

hibited in the subjective probability assessments made by less experienced individuals. Inexperienced forecasters sometimes produce probability forecasts exhibiting overconfidence (Murphy, 1985), or biases due to such factors as excessive reliance on recently acquired information (Spetzler and Staël von Holstein, 1975).

Individuals who are experienced at assessing their subjective probabilities can do so in a seemingly unconscious or automatic manner. People who are new to the practice often find it helpful to use devices that allow comparison of the uncertainty to be assessed with an uncertain situation that is more concrete and familiar. For example, Spetzler and Staël von Holstein (1975) describe a physical device called a *probability wheel*, which consists of a spinner of the sort that might be found in a child's board game, on a background that has the form of a pie chart. This background has two colors, blue and orange, and the proportion of the background covered by each of the colors can be adjusted. The probability wheel is used to assess the probability of a dichotomous event (e.g., a PoP forecast) by adjusting the relative coverages of the two colors until the forecaster feels that the probability of the event to be forecast is about equal to the probability of the spinner stopping in the orange sector. The subjective probability forecast is then read as the angle subtended by the orange section, divided by 360°.

Conceptual devices can also be employed to assess subjective probabilities. For many people, comparison of the uncertainty surrounding the future weather is most easily assessed in the context of lottery games or betting games. Such conceptual devices translate the probability of an event to be forecast into more concrete terms by posing such questions as "Would you prefer to be given \$2 if precipitation occurs tomorrow, or \$1 for sure (regardless of whether precipitation occurs)?" Individuals preferring the sure \$1 in this "lottery" situation evidently feel that the relevant PoP is less than 0.5, while individuals who feel the PoP is greater than 0.5 would prefer to receive \$2 on the chance of precipitation. A forecaster can use this "lottery" device by adjusting the variable payoff relative to the "certainty equivalent" (the sum to be received for sure) to the point where either choice would be equally attractive. That is, the variable payoff is adjusted until the expected (i.e., probability-weighted average—see Chapter 4) payment is equal to the certainty equivalent. With the subjective probability denoted as p, the procedure can be written formally as

$$\begin{aligned} \text{Expected payoff} &= p(\text{variable payoff}) + (1 - p)(\$0) \\ &= \text{certainty equivalent}, \end{aligned} \tag{6.30a}$$

which leads to

$$p = \frac{\text{certainty equivalent}}{\text{variable payoff}}. \tag{6.30b}$$

The same kind of logic can be applied in an imagined betting situation. Here forecasters ask themselves whether receiving a specified payment should the weather event to be forecast occurs, or suffering some other monetary loss if the

event does not occur, is preferable. In this case the subjective probability as assessed by finding monetary amounts for the payment and loss such that the bet is a "fair" one, implying that the forecaster would be equally happy to be on either side of it. Since the expected payoff from a fair bet is zero, the betting game situation can be represented as

$$\text{Expected payoff} = p(\$\ \text{payoff}) + (1 - p)(-\ \$\ \text{loss}) = 0, \tag{6.31a}$$

leading to

$$p = \frac{\$\ \text{loss}}{\$\ \text{loss} + \$\ \text{payoff}}. \tag{6.31b}$$

6.6.4 Assessing Continuous Distributions

The same kinds of lotteries or betting games just described can also be used to assess points on a subjective continuous probability distribution using the *method of successive subdivision.* Here the approach is to identify quantiles of the subjective distribution by comparing the probabilities they imply with the reference probabilities derived from conceptual money games. Use of this method in an operational setting is described in Krzysztofowicz *et al.* (1993).

The easiest quantile to identify is the median. Suppose that the distribution to be identified is for tomorrow's maximum temperature. Since the median divides the subjective distribution into two equally probable halves, its location can be assessed by evaluating one's preference between, say, \$1 for sure and \$2 if tomorrow's maximum temperature is warmer than 14°C. The situation is the same as that described in Eqs. (6.30). Preference for the certainty of \$1 implies a subjective probability for the event {maximum temperature warmer than 14°C} that is smaller than 0.5. A forecaster preferring the chance at \$2 evidently feels that the probability for this event is larger than 0.5. Since the cumulative probability, p, for the median is fixed at 0.5, the threshold defining the event {outcome above median} is adjusted until one is indifferent between the certainty equivalent and a variable payoff equal to twice the certainty equivalent.

The quartiles can be assessed in the same way, except that the ratios of certainty equivalent to variable payoff must correspond to the cumulative probabilities of the quartiles: specifically, 1/4 or 3/4. At what temperature lower-quartile T_{lq} is one indifferent to the alternatives of receiving \$1 for sure, or \$4 if tomorrow's maximum temperature is below T_{lq}? The temperature T_{lq} then estimates the forecaster's subjective lower quartile. Similarly, the upper-quartile temperature T_{uq}, at which one is indifferent to the alternatives of \$1 for sure or \$4 if the temperature is above T_{uq}, estimates the upper quartile.

Especially when one is inexperienced at probability assessments, it is a good idea to perform some consistency checks. In the method just described, the quartiles are assessed independently, but together define a range—the 50% central credible interval—in which half the probability should lie. Therefore a good

check on their consistency would be to verify that one is indifferent to the choices between \$1 for sure, and \$2 if $T_{lq} < T \leq T_{uq}$. If one prefers the certainty equivalent in this comparison, the quartile estimates T_{lq} and T_{uq} are apparently too close. If one prefers the chance at the \$2, they apparently delineate too much probability. Similarly, one could verify indifference between the certainty equivalent, and twice the certainty equivalent if the temperature falls between the median and one of the quartiles. Any inconsistencies discovered in checks of this type indicate that some or all of the previously estimated quantiles need to be reassessed.

Exercises

6.1. Derive a simple linear regression equation using the data in Table A.3, relating June temperature (as the predictand) to June pressure (as the predictor).
   (a) Explain the physical meanings of the two parameters.
   (b) Formally test whether the fitted slope is significantly different from zero.
   (c) Compute the R^2 statistic.
   (d) Estimate the probability that a predicted value corresponding to $x_0 =$ 1013 mb will be within 1°C of the regression line, using Eq. (6.21).
   (e) Repeat (d), assuming the prediction variance equals the MSE.

6.2. Consider the following ANOVA table, describing the results of a regression analysis:

Source	df	SS	MS	F
Total	23	2711.60		
Regression	3	2641.59	880.53	251.55
Residual	20	70.01	3.50	

   (a) How many predictor variables are in the equation?
   (b) What is the sample variance of the predictand?
   (c) What is the R^2 value?
   (d) Estimate the probability that a prediction made by this regression will be within ± 2 units of the actual value.

6.3. The 19 nonmissing precipitation values in Table A.3 can be used to fit the regression equation:

$$\ln[(\text{precipitation}) + 1 \text{ mm}] = 499.4 - 0.512(\text{pressure}) + 0.796(\text{temperature})$$

   The MSE for this regression is 0.701. (The constant 1 mm has been added to ensure that the logarithm is defined for all data values.)
   (a) Estimate the missing precipitation value for 1956 using this equation.
   (b) Construct a 95% confidence interval for the estimated 1956 precipitation.

6.4. Explain how to use cross-validation to estimate the prediction mean-squared error, and the sampling distribution of the regression slope, for the problem in Exercise 6.1. If the appropriate computing resources are available, implement your algorithm.

6.5. Hurricane Zeke is an extremely late storm in a very busy hurricane season. It has recently formed in the Carribean, the 500-mb height at "gridpoint 37" (relative to the storm) is 5400 m, the 500-mb height at "gridpoint 3" is 5500 m, and the 1000-mb height at "gridpoint 51" is −200 m (i.e., the surface pressure near the storm is well below 1000 mb).

(a) Use the NHC 67 model (Table 6.6) to forecast the westward component of its movement over the next 12 h, if storm has moved 80 nautical miles due westward in the last 12 h.

(b) What would the NHC 67 forecast of the westward displacement be if, in the previous 12 h, the storm had moved 80 nautical miles westward *and* 30 nautical miles northward (i.e., P_y = 30 nautical miles)?

6.6. A formerly operational MOS equation for predicting fall (September, October, November) maximum temperature (°F) at Binghamton, New York, at the 60-h projection was

$$\begin{aligned}\text{Max } T = &-363.2 + 1.541(\text{850-mb } T)\\ &- 0.1332(\text{SFC-490 mb RH}) - 10.3(\cos \text{DOY})\end{aligned}$$

where (850-mb T) is the 48-h NWP forecast of temperature (K) at 850 mb; (SFC-490 mb RH) is the 48-h NWP forecast lower tropospheric relative humidity (RH) in percent (%), (cos DOY) is the cosine of the Julian date transformed to radians or degrees, i.e., = $\cos(2\pi t/365)$ or = $\cos(360°t/365)$; and t is the Julian date of the valid time. (The Julian date for January 1 is 1, and for October 31 it is 304.)

Calculate what the 60-h MOS maximum temperature forecast would be for the following:

	Valid time	48-h 850-mb T Forecast [kelvins (K)]	48-h mean RH Forecast (%)
(a)	September 4	278	30
(b)	November 28	278	30
(c)	November 28	258	30
(d)	November 28	278	90

6.7. A MOS equation for 12–24-h PoP in the warm season might look something like:

$$\begin{aligned}\text{PoP} = &\ 0.25 + 0.0063(\text{mean RH}) - 0.163(0\text{–}12 \text{ ppt [bin @ 0.1 in.]})\\ &- 0.165(\text{mean RH [bin @ 70\%]}),\end{aligned}$$

where mean RH is the same variable as in Exercise 6.6 (in %) for the appropriate projection, 0–12 ppt is the NWP-forecast precipitation amount in

the first 12 h of the forecast, and [bin @ xxx] indicates use as a binary variable: =1 if the predictor is ≤xxx, =0 otherwise. Then evaluate the MOS PoP forecasts for the following conditions:

	12-h mean RH (%)	0–12 ppt (inches)
(a)	90	0.00
(b)	65	0.15
(c)	75	0.15
(d)	75	0.09

6.8. Explain why the slopes of the solid lines decrease, from Fig. 6.18 to Fig. 6.19a, to Fig. 6.19b. What would the corresponding MOS equation be for an arbitrarily long projection into the future?

6.9. A forecaster is equally happy with the prospect of receiving $10 for sure, or $50 if freezing temperatures occur on the following night. What is the forecaster's subjective probability for frost?

6.10. A forecaster is indifferent between receiving $1 for sure and any of the following: $8 if tomorrow's rainfall is greater than 55 mm, $4 if tomorrow's rainfall is greater than 32 mm, $2 if tomorrow's rainfall is greater than 12 mm, $1.33 if tomorrow's rainfall is greater than 5 mm, and $1.14 if tomorrow's precipitation is greater than 1 mm.

(a) What is the median of this subjective distribution?

(b) What would be a consistent 50% central credible-interval forecast? A 75% central credible-interval forecast?

(c) In this forecaster's view, what is the probability of receiving between 1 and 32 mm of precipitation?

Chapter 7 Forecast Verification

7.1 Background

Forecast verification is the process of determining the quality of forecasts. Verification of weather forecasts has been undertaken since at least 1884 (Muller, 1944). Perhaps not surprisingly, there can be differing views of what constitutes a good forecast (Murphy, 1993). A wide variety of forecast verification procedures exist, but all involve measures of the relationship between a forecast or set of forecasts, and the corresponding observation(s) of the predictand. Any forecast verification method thus necessarily involves comparisons between matched pairs of forecasts and the observations to which they pertain.

On a fundamental level, forecast verification involves investigation of the properties of the joint distribution of forecasts and observations (Murphy and Winkler, 1987). That is, any given verification data set consists of a collection of forecast/observation pairs whose joint behavior can be characterized in terms of the relative frequencies of the possible combinations of forecast/observation outcomes. A parametric joint distribution such as the bivariate normal (Chapter 4) can sometimes be useful in representing this joint distribution for a particular data set, but the empirical joint distribution of these quantities (more in the spirit of Chapter 3) more usually forms the basis of forecast evaluation measures. Ideally the association between forecasts and the observations to which they pertain will be reasonably strong, and the nature and strength of this association will be reflected in their joint distribution.

7.1.1 Purposes of Forecast Verification

Objective evaluations of forecast quality are undertaken for a variety of reasons. Brier and Allen (1951) categorized these as serving administrative, scientific, and economic purposes. In this view, administrative use of forecast verification pertains to ongoing monitoring of operational forecasts. For example, it is often of interest to examine trends of forecast performance through time. Rates of forecast improvement, if any, for different locations or lead times can be compared. Verification of forecasts from different sources for the same events can also be compared. Here forecast verification techniques allow comparison of the relative

merits of competing forecasters or forecasting systems. This is frequently the purpose of forecast verification in scoring student "forecast games" at colleges and universities.

Analysis of verification statistics and their components can also help in the assessment of specific strengths and weaknesses of forecasters or forecasting systems. Although classified by Brier and Allen as "scientific," this application of forecast verification is perhaps better termed *diagnostic verification* (Murphy *et al.*, 1989; Murphy and Winkler, 1992). Here specific attributes of the relationship between forecasts and the subsequent events are investigated, that can highlight strengths and deficiencies in a set of forecasts. Human forecasters can be given feedback on the performance of their forecasts in different situations that will hopefully lead to better forecasts in the future. Similarly, forecast verification measures can point to problems in forecasts produced by objective means, possibly leading to better forecasts through methodological improvements.

Ultimately, the justification for any forecasting enterprise is that it supports better decision making. The usefulness of forecasts to support decision making clearly depends on their error characteristics, which are elucidated through forecast verification methods. Thus the economic motivations for forecast verification are to provide the information necessary for users to derive full value from forecasts, and to enable estimation of that economic value. However, since the economic value of forecast information in different decision situations must be evaluated on a case-by-case basis (Wilks, 1995), forecast value cannot be computed from the verification statistics alone. Similarly, although it is sometimes possible to guarantee, on the basis of a detailed verification analysis, the economic superiority of one forecast source over another for all forecast users (Ehrendorfer and Murphy, 1988; Krzysztofowicz and Long, 1990, 1991), superiority with respect to a single verification measure does not always imply superior forecast value for all users.

7.1.2 Factorization of the Joint Distribution of Forecasts and Observations

Of fundamental interest with respect to the verification of forecasts is the joint distribution of the forecasts and observations. In practical settings, both the forecasts and observations are discrete variables. That is, if the forecasts and observations are not already discrete quantities, they are rounded operationally to one of a finite set of values. Denote the forecast by y_i, which can take on any of the I values $y_1, y_2, \ldots, y_I$; and the corresponding observation as o_j, which can take on any of the J values $o_1, o_2, \ldots, o_J$. Then the joint distribution of the forecasts and observations is denoted

$$p(y_i, o_j) = \Pr\{y_i, o_j\} = \Pr\{y_i \cap o_j\}; \qquad i = 1, \ldots, I;\ j = 1, \ldots, J. \tag{7.1}$$

This is a discrete bivariate probability distribution function, associating a probability with each of the $I \times J$ possible combinations of forecasts and observations.

Even in the simplest cases, for which $I = J = 2$, this joint distribution can be difficult to use directly. Using the definition of conditional probability [Eq. (2.9)], the joint distribution can be factored in two ways that are informative about different aspects of the verification problem. From a forecasting standpoint, the more familiar and intuitive of the two is

$$p(y_i, o_j) = p(o_j \mid y_i)p(y_i); \qquad i = 1, \ldots, I; j = 1, \ldots, J, \quad (7.2)$$

which has been called the *calibration-refinement* factorization (Murphy and Winkler, 1987). One part of this factorization consists of a set of I conditional distributions, $p(o_j \mid y_i)$, each of which consists of probabilities for all the J outcomes o_j, given one of the forecasts y_i. That is, this conditional distribution specifies how often each possible weather event occurred on those occasions when the single forecast y_i was issued, or how well each forecast y_i is *calibrated.* The other part of this factorization is the unconditional (*marginal*) distribution $p(y_i)$, which specifies the relative frequencies of use of each of the forecast values y_i, or how often each of the I possible forecast values is used. This marginal distribution is sometimes called the *predictive distribution* of the forecasts.

The other possible factorization of the joint distribution of forecasts and observations is the *likelihood-base rate* factorization (Murphy and Winkler, 1987),

$$p(y_i, o_j) = p(y_i \mid o_j)p(o_j); \qquad i = 1, \ldots, I; j = 1, \ldots, J. \quad (7.3)$$

Here the conditional distribution $p(y_i \mid o_j)$ expresses the likelihoods that each of the allowable forecast values y_i would have been issued in advance of the observed weather event o_j. Although this concept may seem logically reversed, it can reveal useful information about the nature of forecast performance. The unconditional distribution $p(o_j)$ consists simply of the relative frequencies of the J weather events o_j in the verification data set, and is usually called the *sample climatology.*

7.1.3 Scalar Attributes of Forecast Performance

Although the two factorizations of the joint distribution of forecasts and observations can help organize the verification information conceptually, neither reduces the *dimensionality* (Murphy, 1991), or degrees of freedom, of the verification problem. That is, since all the probabilities in the joint distribution [Eq. (7.1)] must add to 1, it is completely specified by any $(I \times J) - 1$ of these probabilities. The factorizations of Eqs. (7.2) and (7.3) reexpress this information differently and informatively, but $(I \times J) - 1$ distinct probabilities are still required to completely specify each factorization.

Even in the simplest case of $I = J = 2$, therefore, complete specification of forecast performance requires a three-dimensional verification measure. This level of dimensionality is already sufficient to make understanding and comparison of forecast evaluation statistics difficult. The difficulty is compounded in the many verification situations where $I > 2$ and/or $J > 2$. As a consequence, it is tradi-

tional to summarize forecast performance using one or several scalar (i.e., one-dimensional) verification measures. While many of the scalar summary statistics have been found through analysis and experience to provide very useful information about forecast performance, some of the information in the full joint distribution of forecasts and observations is inevitably discarded when the dimensionality of the verification problem is reduced.

The following is a list of scalar aspects, or attributes, of forecast quality. Many of them are interrelated.

1. *Accuracy* refers to the average correspondence between individual forecasts and the events they predict. Scalar measures of accuracy are meant to summarize, in a single number, the overall quality of a set of forecasts. This concept may seem straightforward and unambiguous, but many possible measures of forecast accuracy are potentially applicable to a given forecast format. Several of the more common measures of accuracy will be presented in subsequent sections. The remaining forecast attributes in this list can often be interpreted as components, or aspects, of accuracy.

2. *Bias*, or unconditional bias, or systematic bias, measures the correspondence between the average forecast and the average observed value of the predictand. This is different from accuracy, which measures the average correspondence between individual pairs of forecasts and observations. Temperature forecasts that are consistently too warm or precipitation forecasts that are consistently too wet both exhibit bias, regardless of whether the forecasts are reasonably accurate or quite inaccurate.

3. *Reliability*, or calibration, or conditional bias, pertains to the relationship of the forecast to the average observation for specific values of the forecast. Reliability measures sort the forecast/observation pairs into groups according to the value of the forecast variable, and characterize the conditional distributions of the observations given the forecasts. Thus, measures of reliability summarize the J conditional distributions $p(o_i \mid y_j)$ of the calibration-refinement factorization [Eq. (7.2)].

4. *Resolution* refers to the degree to which the forecasts sort the observed events into groups that are different from each other. It is related to reliability, in that both are concerned with the properties of the conditional distributions of the observations given the forecasts, $p(o_i \mid y_j)$. Therefore, resolution also relates to the calibration-refinement factorization of the joint distribution of forecasts and observations. However, *resolution* pertains to the differences between the conditional averages of the observations for different values of the forecast, whereas *reliability* compares the conditional averages of the observations with the forecast values themselves. If average temperature outcomes following forecasts of, say, 10 and 20° C are very different, the forecasts can resolve these different temperature outcomes, and are said to exhibit resolution. If the temperature outcomes following forecasts of 10 and 20° C are nearly the same on average, the forecasts exhibit almost no resolution.

5. *Discrimination* is the converse of resolution. Measures of discrimination summarize the conditional distributions of the forecasts given the observations, $p(y_j \mid o_i)$, in the likelihood-base rate factorization [Eq. (7.3)] of the joint distribution of forecasts and observations. The discrimination attribute reflects the ability of the forecasting system to produce different forecasts for those occasions having different realized outcomes of the predictand. If a forecasting system forecasts f = snow with equal frequency when o = snow and o = sleet, the two conditional probabilities of a forecast of snow are equal, and the forecasts are not able to discriminate between snow and sleet events.

6. *Sharpness*, or refinement, is an attribute of the forecasts alone, without regard to their corresponding observations. Measures of sharpness characterize the unconditional distribution (relative frequencies of use) of the forecasts, $p(y_j)$ in the calibration-refinement factorization [Eq. (7.2)]. Forecasts that rarely deviate much from the climatological value of the predictand exhibit low sharpness. In the extreme, forecasts consisting only of the climatological value of the predictand exhibit no sharpness. By contrast, forecasts that are frequently much different from the climatological value of the predictand are sharp. Sharp forecasts exhibit the tendency to "stick their necks out." Sharp forecasts will be accurate only if they are also well calibrated; anyone can produce sharp forecasts, but the difficult task is to ensure that these forecasts correspond well to the subsequent observations.

7.1.4 Forecast Skill

Forecast skill refers to the relative accuracy of a set of forecasts, with respect to some set of standard *control*, or *reference*, forecasts. Common choices for the reference forecasts are climatological average values of the predictand, persistence forecasts (values of the predictand in the previous time period), or random forecasts (with respect to the climatological relative frequencies of the forecast events, o_j). Other choices for the reference forecasts can be more appropriate in some cases. For example, when evaluating the performance of a new forecasting system, it might be appropriate to compute skill relative to the forecasts that this new system might replace.

Forecast skill is usually presented as a *skill score* (SS), which is interpreted as a percentage improvement over the reference forecasts. In generic form, the skill score for forecasts characterized by a particular measure of accuracy A, with respect to the accuracy A_{ref} of a set of reference forecasts, is given by

$$\text{SS}_{\text{ref}} = \frac{A - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} \times 100\%, \tag{7.4}$$

where A_{perf} is the value of the accuracy measure that would be achieved by perfect forecasts. Note that the generic skill score formulation gives consistent results whether the accuracy measure has a positive (larger values of A are better) or

negative (smaller values of A are better) orientation. If $A = A_{perf}$, the skill score attains its maximum value of 100%. If $A = A_{ref}$, then $SS_{ref} = 0\%$, indicating no improvement over the reference forecasts. If the forecasts being evaluated are inferior to the reference forecasts, $SS_{ref} < 0\%$.

7.2 Categorical Forecasts of Discrete Predictands

Forecast verification is perhaps easiest to understand with reference to categorical forecasts of discrete predictands. *Categorical* means that the forecast consists of a flat statement that one and only one of a set of possible events will occur. Categorical forecasts contain no expression of uncertainty, in distinction to probabilistic forecasts. A *discrete* predictand is an observable variable that takes on one and only one of a finite set of possible values. This is in distinction to a continuous predictand, which (at least conceptually) may take on any value on the relevant portion of the real line.

In this verification setting there is typically a one-to-one correspondence between the allowable forecast values and the observable predictand values. In terms of the joint distribution of forecasts and observations, $I = J$. The simplest possible situation is for the dichotomous $I = J = 2$ case, or verification of categorical "yes/no" forecasts. Here there are $I = 2$ possible forecasts, either that the event will ($i = 1$, or y_1) or will not ($i = 2$, or y_2) occur. Similarly, there are $J = 2$ possible outcomes: Either the event subsequently occurs (o_1) or it does not (o_2).

7.2.1 The Contingency Table

Conventionally, categorical verification data is displayed in an $I \times J$ *contingency table* of absolute frequencies, or counts, of the $I \times J$ possible combinations of forecast and event pairs. If these counts are transformed to relative frequencies, by dividing each tabulated entry by the sample size (total number of forecast/event pairs), the (sample) joint distribution of forecasts and observations [Eq. (7.1)] is obtained. Figure 7.1 illustrates the essential equivalence of the contingency table and the joint distribution of forecasts and observations for the simple, $I = J = 2$, case. The boldface portion of Fig. 7.1a shows the arrangement of the four possible combinations of forecast/event pairs as a contingency table. For example, the event in question was successfully forecast to occur a times out of n total forecasts. Similarly, on b occasions the event was forecast to occur but did not. The total number of forecast/event pairs in the data set is $n = a + b + c + d$.

The joint distribution of these forecasts and observations, $p(y_i, o_j)$, consists simply of the relative frequencies with which each of the possible forecast/event pairs occurred, and is obtained from the contingency table of raw counts (a–d) by dividing through by the sample size. The boldface portion of Fig. 7.1b shows this discrete, bivariate distribution. To relate the representation of this joint distribu-

Fig. 7.1 Relationship between counts (letters a–d) of forecast/event pairs for the dichotomous categorical verification situation as displayed in a 2×2 contingency table (bold, panel a), and the corresponding joint distribution of forecasts and observations $[p(y, o)]$ (bold, panel b). Also shown are the marginal totals, indicating how often both events were forecast and observed in absolute terms; and the marginal distributions of the observations $[p(o)]$ and forecasts $[p(y)]$, which indicates the same information in relative frequency terms.

tion in Fig. 7.1b to the subscripted notation, define $i = 1$ and $j = 1$ as "yes," and $i = 2$ and $j = 2$ as "no." Then, for example, the relative frequency of a successful "yes" forecast would be $p(y_1, o_1) = a/n$.

It is also common to include what are called *marginal totals* with a contingency table of counts. These are simply the row and column totals yielding, in this case, the numbers of times each "yes" or "no" forecast or observation, respectively, occurred. These are shown in Fig. 7.1a in normal typeface, as is the sample size, $n = a + b + c + d$. Expressing the marginal totals in relative frequency terms, again by dividing through by the sample size, yields the marginal distribution of the forecasts, $p(y_i)$, and the marginal distribution of the observations, $p(o_j)$. These two distributions appear in Eqs. (7.2) and (7.3), respectively. Similarly, the conditional distributions $p(o_j \mid y_i)$ and $p(y_j \mid o_i)$ in the same two equations can be obtained by considering only relative frequencies *within* rows or columns. For example, the two probabilities representing the conditional distribution of the observations given a "yes" forecast, $p(o_j \mid y_1)$, are computed from the entries in the first row by applying Eq. (2.9). These two probabilities are $p(o_1 \mid y_1) = [a/n]/[(a + b)/n] = a/(a + b)$, and $p(o_2 \mid y_1) = [b/n]/[(a + b)/n] = b/(a + b)$.

7.2.2 Accuracy Measures for Binary Forecasts

Perfectly accurate forecasts in the 2×2 categorical forecasting situation will clearly exhibit $b = c = 0$, with all "yes" forecasts for the event followed by the

event and all "no" forecasts for the event followed by nonoccurrence. For real, imperfect forecasts, measures are needed to discern degrees of accuracy. Several scalar measures are in common use, with each reflecting somewhat different aspects of the underlying joint distribution.

Perhaps the most direct and intuitive measure of the accuracy of categorical forecasts is the *hit rate*, or the *proportion correct*. This is simply the fraction of the n forecasting occasions when the categorical forecast correctly anticipated the subsequent event or nonevent. In terms of Fig. 7.1a, the hit rate is given by

$$H = \frac{a + d}{n}. \tag{7.5}$$

The hit rate satisfies the principle of *equivalence of events*, since it credits correct "yes" and "no" forecasts equally. As Example 7.1 will show, however, this is not always a desirable attribute. The hit rate also penalizes both kinds of errors equally. The worst possible hit rate is zero. The best possible hit rate is one. Sometimes H in Eq. (7.5) is multiplied by 100%, and referred to as the *percent correct*, or the *percentage of forecasts correct* (PFC).

A frequently used alternative to the hit rate, particularly when the event to be forecast (as the "yes" event) occurs substantially less frequently than the nonoccurrence ("no"), is the *threat score* (TS), or the *critical success index* (CSI). In terms of Fig. 7.1a, the threat score is computed as

$$\text{TS} = \text{CSI} = \frac{a}{a + b + c}. \tag{7.6}$$

The threat score is the number of correct "yes" forecasts divided by the total number of occasions on which that event was forecast and/or observed. It can be viewed as a hit rate for the quantity being forecast, after removing correct "no" forecasts from consideration. The worst possible threat score is zero, and the best possible threat score is one.

Two other measures are also routinely computed for 2×2 categorical verification data. The first is called the *probability of detection* (POD), which is referred to in the older literature as *prefigurance*. The POD is simply the fraction of those occasions when the forecast event occurred on which it was also forecast. That is, the POD is the likelihood that the event would be forecast, given that it occurred, or $p(y_1 \mid o_1)$ in the notation of Eq. (7.3). In terms of the entries in Fig. 7.1a, the probability of detection is computed as

$$\text{POD} = \frac{a}{a + c}. \tag{7.7}$$

The POD for perfect forecasts is one, and the worst POD is zero.

The *false-alarm rate* (FAR) is that proportion of forecast events that fail to materialize. This is an element of the conditional distribution of events given the

forecast, specifically $p(o_2 \mid y_1)$. In terms of Fig. 7.1, the FAR is computed as

$$\text{FAR} = \frac{b}{a + b}. \tag{7.8}$$

The FAR has a negative orientation, so that smaller values of FAR are to be preferred. The best possible FAR is zero, and the worst possible FAR is one. The FAR is the complement of a quantity called the *postagreement* in the older literature. That is, FAR = 1 − postagreement. The postagreement is $a/(a + b) = p(o_1 \mid y_1)$.

7.2.3 Bias

The *bias*, or comparison of the average forecast with the average observation, of categorical forecasts, is usually represented as a ratio. In terms of Fig. 7.1a the bias ratio is

$$B = \frac{a + b}{a + c}. \tag{7.9}$$

The bias is simply the ratio of the number of "yes" forecasts to the number of "yes" observations. Unbiased forecasts exhibit $B = 1$, indicating that the event was forecast the same number of times that it was observed. Note that this says nothing about the correspondence between the forecasts and observations of the event on particular occasions, so that the bias is not an accuracy measure. Bias greater than one indicates that the event was forecast more often than observed, which is called *overforecasting*. Conversely, bias less than one indicates that the event was forecast less often than observed, or was *underforecast*.

Example 7.1. The Finley Tornado Forecasts

A historical 2 × 2 categorical forecast data set that is often used to illustrate evaluation of forecasts in this format is the collection of Finley's tornado forecasts (Finley, 1884). John Finley was a sergeant in the U.S. Army who, using telegraphed synoptic information, formulated yes/no tornado forecasts for 18 regions of the United States east of the Rocky Mountains. The contingency table for his $n = 2803$ forecasts is presented in Table 7.1a.

Finley chose to evaluate his forecasts using the hit rate [Eq. (7.5)], which for his data is $H = (28 + 2680)/2803 = 0.966$. On the basis of the hit rate, Finley claimed 96.6% accuracy. However, the hit rate for this data set is dominated by the correct "no" forecasts, since tornados are relatively rare. Shortly after Finley's paper appeared, a critic pointed out that always forecasting "no" would produce an even higher hit rate. The contingency table that would be obtained if tornados were never forecast is shown in Table 7.1b. These hypothetical forecasts yield a hit rate of $H = (0 + 2752)/2803 = 0.982$, which is, indeed, higher than the hit rate for the actual forecasts.

Table 7.1

Contingency Tables for Verification of the Finley Tornado Forecasts, from 1884[a]

a. Original Finley Forecasts

		Tornados observed	
		Yes	No
Tornados forecast	Yes	28	72
	No	23	2680

b. Finley Data with "No Tornados" Always Forecast

		Tornados observed	
		Yes	No
Tornados forecast	Yes	0	0
	No	51	2752

[a] The forecast event is occurrence of a tornado, with separate forecasts for 18 regions of the United States east of the Rocky Mountains. (a) The table for the forecasts as originally issued; (b) data that would have been obtained if "no tornados" had always been forecast (in both tables, $n = 2803$).

Employing the threat score gives a more reasonable comparison. For Finley's original forecasts, the threat score is TS = 28/(28 + 72 + 23) = 0.228, whereas for the obviously useless "no" forecasts in Table 7.1b the threat score is TS = 0/(0 + 0 + 51) = 0. Clearly the threat score would be preferable to the hit rate in this instance, but it is still not completely satisfactory. Equally useless would be a forecasting system that always forecast "yes" for tornados. For constant "yes" forecasts the threat score would be TS = 51/(51 + 2752 + 0) = 0.018, which is small, but not zero.

The POD for Finley's tornado forecasts is POD = 28/(28 + 23) = 0.549. The FAR for these forecasts is FAR = 72/(28 + 72) = 0.720. The Finley tornado forecasts in Table 7.1a exhibit $B = (28 + 72)/(28 + 23) = 1.96$, indicating that Finley overforecast the tornados by a factor of almost 2. ¤

7.2.4 Accuracy Measures for Multicategory Forecasts

Categorical forecasts are not limited to the 2 × 2 format, although that simple situation is the easiest to understand. In some settings it is natural or desirable to consider and forecast more than two MECE events. The left side of Fig. 7.2, in boldface type, shows a generic contingency table for the case of $I = J = 3$ possible forecasts and events. Here the counts for each of the nine possible forecast/event pair outcomes are denoted by the letters r through z, yielding a total sample size $n = r + s + t + u + v + w + x + y + z$. As before, dividing each of the nine counts in this 3 × 3 contingency table by the sample size would yield the joint distribution of forecasts and observations, $p(y_i, o_j)$.

Of the accuracy measures listed above, only the hit rate [Eq. (7.5)] generalizes directly to situations with more than two forecast and event categories. Regardless

Fig. 7.2 Contingency table for the $I = J = 3$ categorical forecast verification situation (bold), and its reduction to three 2×2 contingency tables. Each 2×2 contingency table is constructed by regarding one of the three original events as "the" event being forecast, and the remaining two original events combined as complementary, i.e., "not the forecast event." For example, the 2×2 table for "event 1" lumps event 2 and event 3 as the single event, "not event 1." The letters a, b, c, and d are used in the same sense as in Fig. 7.1a. Performance measures specific to the 2×2 contingency tables, such as the threat score, can then be computed separately for each of the resulting tables. This procedure generalizes easily to square forecast verification contingency tables with arbitrarily many forecast and event categories.

of the size of I and J, the hit rate is still given by the proportion of correct forecasts, or "hits," in the sample of n forecasts. The number of hits is obtained by adding the counts along the diagonal from the upper left to the lower right corners of the contingency table. In Fig. 7.2, the numbers r, v, and z represent the numbers of occasions when the first, second, and third events were correctly forecast, respectively. Therefore, in the 3×3 table represented in this figure, the hit rate would be $H = (r + v + z)/n$.

The bias ratio [Eq. (7.9)] also generalizes to the case of $I = J > 2$ forecasts and events. For the 3×3 contingency table in Fig. 7.2, the bias for forecasts of event 1 would be $B_1 = (r + s + t)/(r + u + x)$, the bias for forecasts of event 2 would be $B_2 = (u + v + w)/(s + v + y)$, and the bias for forecasts of event 3 would be $B_3 = (x + y + z)/(t + w + z)$.

The other verification measures presented in the previous sections pertain only to the dichotomous yes/no forecast situation. In order to apply these to categorical forecasts that are not dichotomous, it is necessary to collapse the $I = J > 2$ contingency table into a series of 2×2 contingency tables. Each of these 2×2 tables is constructed, as indicated in Fig. 7.2, by considering "the forecast event" in distinction to the complementary "not the forecast event." This complementary event is simply constructed as the union of the $J - 1$ remaining events. In Fig. 7.2, the 2×2 contingency table for "event 1" lumps events 2 and 3 as "not event 1." Thus, the number of times event 1 is correctly forecast is still $a = r$, and the number of times it is incorrectly forecast is $b = s + t$. From the standpoint of this collapsed 2×2 contingency table, whether the incorrect forecast of event 1 was followed by event 2 or event 3 is unimportant. Similarly, the number of times the

event "not event 1" is correctly forecast is $d = v + w + y + z$, and includes cases where event 2 was forecast but event 3 occurred, and event 3 was forecast but event 2 occurred.

Example 7.2. A Set of Multicategory Forecasts

The threat score, POD, FAR, and bias are computed separately for each of the derived 2×2 contingency tables. The data in Table 7.2 illustrate the procedure. On the left of this table is a 3×3 verification contingency table for forecasts of freezing rain (y_1), snow (y_2), and rain (y_3) from Goldsmith (1990). These are categorical forecasts, conditional on the occurrence of some form of precipitation, for the eastern region of the United States, for the months of October through March of 1983/84 through 1988/89. They are MOS forecasts of the form of the "PTYPE" forecasts in Table 6.7, but produced by an earlier MOS system. For each of the three precipitation types, a 2×2 contingency table can be constructed, following Fig. 7.2, that summarizes the performance of forecasts of that precipitation type in distinction to the other two precipitation types together. The summary measures computed from the 2×2 contingency tables are reasonably consistent with each other, and indicate that the rain forecasts were slightly superior to the snow forecasts, but that the freezing rain forecasts were substantially less successful. ¤

7.2.5 Conversion of Probabilistic to Categorical Forecasts

The MOS systems from which the data in Table 7.2 and Table 6.7 were taken actually produce probability forecasts for precipitation type, as well as for some other variables whose forecasts are disseminated in categorical format. That is, the MOS equations underlying the categorical forecasts summarized in Table 7.2 actually specified probabilities for each of the three precipitation type outcomes on each forecasting occasion. The publicly issued categorical forecasts were then derived by converting the underlying probabilities to the categorical format. This unfortunate procedure is advocated and practiced under the rationale that categorical forecasts are easier to understand. However, conversion from probability to categorical information inevitably results in a loss of information, to the detriment of the forecast users.

Fundamentally, the conversion from probabilistic to categorical format requires selection of a threshold probability, above which the categorical forecast will be "yes," and below which the categorical forecast will be "no." This procedure seems simple enough; however, the proper threshold to choose depends on the user of the forecast and the particular decision problem(s) to which that user will apply the forecast. Naturally, different decision problems will require different threshold probabilities, and this is the crux of the information-loss issue. In a sense, this conversion from probability to categorical forecasts amounts to the forecaster making the decisions for the forecast users, but without knowing the

Table 7.2

Categorical MOS Forecasts for Freezing Rain (y_1), Snow (y_2), and Rain (y_3)[a]

Full 3 × 3 contingency table				Freezing rain			Snow			Rain		
	o_1	o_2	o_3		o_1	Not o_1		o_2	Not o_2		o_3	Not o_3
y_1	50	91	71	y_1	50	162	y_2	2364	217	y_3	3288	259
y_2	47	2364	170	Not y_1	101	6027	Not y_2	296	3463	Not y_3	241	2552
y_3	54	205	3288									
					TS = 0.160			TS = 0.822			TS = 0.868	
					POD = 0.331			POD = 0.889			POD = 0.932	
					FAR = 0.764			FAR = 0.084			FAR = 0.073	
					B = 1.40			B = 0.97			B = 1.01	

[a] Conditional on occurrence of some form of precipitation, for the eastern region of the United States during cool seasons of 1983/84 through 1988/89. The verification data are presented as a 3 × 3 contingency table on the left, and then as three 2 × 2 contingency tables for each of the three precipitation types. Also shown are the threat score, probability of detection, false alarm rate, and bias ratio for forecasts of each of the three preciptation types. The sample size is n = 6340. [From Goldsmith (1990).]

particulars of the decision problems. Necessarily, then, the conversion from a probability to a categorical forecast is arbitrary.

Example 7.3. Effects of Different Probability Thresholds on Conversion to Categorical Forecasts

It is instructive to examine the procedures used to convert probabilistic to categorical forecasts. Table 7.3 contains a verification data set of probability of precipitation forecasts, issued for the United States during the period October 1980 through March 1981. Here the joint distribution of the $I = 12$ possible forecasts and the $J = 2$ possible observations are presented in the form of the calibration-refinement factorization [Eq. (7.2)]. For each allowable forecast probability y_i, the conditional probability $p(o_1 \mid y_i)$ indicates the relative frequency of the event $j = 1$ (precipitation occurrence) for these $n = 12{,}402$ forecasts. The marginal probabilities $p(y_i)$ indicate the relative frequencies with which each of the $I = 12$ possible forecast values were used.

These precipitation occurrence forecasts were issued as probabilities. If it had been intended to convert them first to a categorical rain/no rain format, a threshold probability would have been chosen in advance. There are several possibilities for this choice, each of which give different results. The two simplest approaches are used rarely, if ever, in operational practice. The first procedure is to forecast the more likely event, which corresponds to selecting a threshold probability of 0.50. The other simple approach is to use the climatological relative frequency of the event being forecast as the threshold probability. For the data set in Table 7.3 this relative frequency is $\Sigma_i p(o_j \mid y_i) p(y_i) = 0.162$ [cf. Eq. (2.13)], although in practice this probability threshold would need to have been estimated in advance using historical climatological data.

The two methods for choosing the threshold probability that are most often used operationally are based on the threat score and the bias, both of which are performance measures for 2×2 contingency tables presented above. For each possible

Table 7.3

Verification Data for Subjective 12–24-h Projection Probability-of-Precipitation Forecasts for United States during October 1980–March 1981[a]

y_i	0.00	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(o_1 \mid y_i)$	.006	.019	.059	.150	.277	.377	.511	.587	.723	.799	.934	.933
$p(y_i)$	.4112	.0671	.1833	.0986	.0616	.0366	.0303	.0275	.0245	.0220	.0170	.0203

[a] Expressed in the form of the calibration-refinement factorization [Eq. (7.2)] of the joint distribution of these forecasts and observations. There are $I = 12$ allowable values for the forecast probabilities, y_i, and $J = 2$ events ($j = 1$ for precipitation and $j = 2$ for no precipitation). The sample climatological relative frequency is 0.162, and the sample size is $n = 12{,}402$.

Fig. 7.3 Derivation of candidate threshold probabilities for converting the probability-of-precipitation forecasts in Table 7.3 to categorical "rain"/"no rain" forecasts. The "Clim" threshold indicates a forecast of rain if the probability is higher than the climatological probability of precipitation, TS_{max} is the threshold that would maximize the threat score of the resulting categorical forecasts, the $B = 1$ threshold would produce unbiased categorical forecasts, and the P_{max} threshold would produce categorical forecasts of the more likely of the two events. Also shown (lighter lines) are the POD and FAR for the resulting 2×2 contingency tables.

choice of a threshold probability, a different 2×2 contingency table, in the form of Fig. 7.1a, results, and therefore different values of TS and B are obtained. When using the threat score to choose the threshold, one selects that threshold producing the maximum threat score. When using the bias ratio, one chooses that threshold producing, as nearly as possible, no bias ($B = 1$).

Figure 7.3 illustrates the dependence of the bias ratio and threat score on the threshold probability for the data given in Table 7.3. Also shown are the POD and FAR values that would be obtained. The threshold probabilities that would be chosen according to the climatological relative frequency (Clim), the maximum threat score (TS_{max}), unbiased categorical forecasts ($B = 1$), and maximum probability (p_{max}) are indicated by the arrows at the bottom of the figure. For example, choosing the relative frequency of precipitation occurrence, 0.162, as the threshold results in forecasts of PoP = 0.00, 0.05, and 0.10 being converted to "no rain," and the other forecasts being converted to "rain." This would have resulted

in $n[p(y_1) + p(y_2) + p(y_3)] = 12{,}402[0.4112 + 0.0671 + 0.1833] = 8205$ "no" forecasts, and $12{,}402 - 8205 = 4197$ "yes" forecasts. Of the 8205 "no" forecasts, one can compute, using the multiplicative law of probability [Eq. (2.10)] that the proportion of occasions that "no" was forecast but precipitation occurred was $p(o_1 \mid y_1)p(y_1) + p(o_1 \mid y_2)p(y_2) + p(o_1 \mid y_3)p(y_3) = (.006)(.4112) + (.019)(.0671) + (.059)(.1833) = 0.0146$. This relative frequency is c/n in Fig. 7.1, so that $c = (0.0146)(12{,}402) = 181$, and $d = 8205 - 181 = 8024$. Similarly, one can compute that, for this cutoff, $a = 12{,}402[(0.150)(0.0986) + \cdots + (0.933)(0.203)] = 1828$ and $b = 2369$. The resulting contingency table yields $B = 2.09$, and TS $= 0.417$. By contrast, the threshold maximizing the threat score is near 0.35, which would have resulted in some overforecasting of precipitation occurrence. ¤

7.2.6 Skill Measures for Contingency Tables

Commonly, forecast verification data in contingency tables are characterized using relative accuracy measures, or skill scores. Historically, a large number of such skill scores have been developed, and many of these are presented by Muller (1944), Murphy and Daan (1985), Stanski *et al.* (1989), and Woodcock (1976). In general the different skill scores perform differently, and sometimes inconsistently. This situation can be disconcerting if one hopes to choose among alternative skill scores, but should not really be surprising given that all of these skill scores are scalar measures of forecast performance in what is intrinsically a higher-dimensional setting. Scalar skill scores are used because they are convenient, but they are necessarily incomplete representations of forecast performance.

The most commonly used skill score for summarizing square contingency tables is the Heidke skill score (Heidke, 1926). The Heidke score is based on the hit rate as the basic accuracy measure, and follows the form of the generic skill score, shown in Eq. (7.4). Thus, perfect forecasts receive Heidke scores of one, forecasts equivalent to the reference forecasts receive zero scores, and forecasts worse than the reference forecasts receive negative scores.

The reference accuracy measure in the Heidke score is the hit rate that would be achieved by random forecasts, subject to the constraint that the marginal distributions of forecasts and observations characterizing the contingency table for the random forecasts, $p(y_i)$ and $p(o_j)$, are the same as the marginal distributions in the actual verification data set. For example, in the 2×2 situation, the marginal probability of a "yes" forecast is $p(y_1) = (a + b)/n$, and the marginal probability of a "yes" observation is $p(o_1) = (a + c)/n$. Since random forecasting implies statistical independence between forecasts and events, the probability of a correct forecast by chance [Eq. (2.11)] is $p(y_1, o_1) = p(y_1)p(o_1) = [(a + b)/n][(a + c)/n] = (a + b)(a + c)/n^2$. Similarly, the probability of a correct "no" forecast by chance is $p(y_2, o_2) = (b + d)(c + d)/n^2$. Thus, for the $I = J = 2$

situation (Fig. 7.1a), the Heidke skill score is

$$\text{HSS} = \frac{(a + d)/n - [(a + b)(a + c) + (b + d)(c + d)]/n^2}{1 - [(a + b)(a + c) + (b + d)(c + d)]/n^2}$$

$$= \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}, \tag{7.10}$$

where the second equality is equivalent but easier to compute. For the Finley tornado forecasts (Table 7.1), HSS = 0.36, indicating that Finley's hit rate of $H = 0.966$ exhibited some skill over random tornado forecasts issued with the same relative frequencies.

The Heidke score can be extended easily to categorical verification problems where there are more than $I = J = 2$ possible forecasts and events. The formula for the Heidke score in this more general case can be most easily written in terms of the joint distribution of forecasts and observations, $p(y_i, o_j)$; the marginal distributions of the forecasts, $p(y_i)$; and the marginal distribution of observations, $p(o_j)$:

$$\text{HSS} = \frac{\sum_{i=1}^{I} p(y_i, o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{i=1}^{I} p(y_i)p(o_i)}. \tag{7.11}$$

Equation (7.11) reduces to Eq. (7.10) for the 2×2 situation.

Another popular skill score for categorical forecasts has been rediscovered many times since being first proposed by Peirce (1884). It is most commonly referred to as the *Hanssen–Kuipers discriminant* (Hanssen and Kuipers, 1965) or *Kuipers' performance index* (Murphy and Daan, 1985), but is sometimes called the *true skill statistic* (TSS) (Flueck, 1987). *Gringorten's* (1967) *skill score* contains equivalent information, as it is a linear transformation of the Kuipers score.

The Kuipers skill score is formulated similarly to the Heidke score, except that the reference hit rate in the denominator is that for random forecasts that are constrained to be unbiased. That is, the imagined random reference forecasts in the denominator have a marginal distribution that is equal to the (sample) climatology, so that $p(y_1) = p(o_1)$ and $p(y_2) = p(o_2)$. In the 2×2 situation of Fig. 7.1, the Kuipers skill score is computed as

$$\text{KSS} = \frac{(a + d)/n - [(a + b)(a + c) + (b + d)(c + d)]/n^2}{1 - [(a + c)^2 + (b + d)^2]/n^2}$$

$$= \frac{ad - bc}{(a + c)(b + d)}, \tag{7.12}$$

where again the second equality is computationally more convenient. Perfect forecasts receive a score of one, random forecasts receive a score of zero, and forecasts inferior to the random forecasts receive negative scores. The Kuipers score can also be extended to verification situations with more than two possible events and

categorical forecasts, using

$$\text{KSS} = \frac{\Sigma_{i=1}^{I}\, p(y_i, o_i) - \Sigma_{i=1}^{I}\, p(y_i)p(o_i)}{1 - \Sigma_{j=1}^{J}\, [p(o_j)]^2}. \tag{7.13}$$

This equation is similar to Eq. (7.11), except that the denominator involves the hit rate for random forecasts having the same relative frequencies as the observations $p(o_j)$. Equation (7.13) reduces to Eq. (7.12) in the 2×2 case of $I = J = 2$.

The Kuipers score possesses a number of appealing characteristics. Both random forecasts and constant forecasts (e.g., always forecasting "no tornados") receive the same (zero) score. Furthermore, unlike the Heidke score, the contribution made to the Kuipers score by a correct "no" or "yes" forecast increases as the event is more or less likely, respectively. Thus, forecasters are not discouraged from forecasting rare events on the basis of their low climatological probability alone. These attributes are characteristic of what has recently been called *equitability* (Gandin and Murphy, 1992), and acceptance of the equitability constraints as desirable leads one to the use of the Kuipers score in 2×2 categorical forecast verification situations. For $I = J > 2$, many skill scores exist that satisfy the equitability constraints. Gerrity (1992) has suggested a family of such scores.

Example 7.4. Heidke and Kuipers' Scores for a 3 × 3 Contingency Table

Using Eq. (7.11), the Heidke score for the 3×3 contingency table in Table 7.2 would be computed as follows. The hit rate would be $H = \Sigma_i p(y_i, o_i) = (50/6340) + (2364/6340) + (3288/6340) = 0.8994$. The hit rate for the random reference forecasts would be $\Sigma_i p(y_i)p(o_i) = (.0334)(.0238) + (.4071)(.4196) + (.5595)(.5566) = 0.4830$. Here, for example, the marginal probability $p(y_1) = (50 + 91 + 71)/6340 = 0.0344$. The hit rate for perfect forecasts is of course one, yielding HSS $= (.8944 - .4830)/(1 - .4830) = 0.8054$.

The computation for the Kuipers score, Eq. (7.13), is the same except that a different reference hit rate is used in the denominator only. This is the unbiased random hit rate $\Sigma_i [p(o_i)^2] = .0238^2 + .4196^2 + .5566^2 = 0.4864$. The Kuipers skill score for this 3×3 contingency table is then KSS $= (.8944 - .4830)/(1 - .4864) = 0.8108$. The difference between the HSS and the KSS for this data is small, because the forecasts exhibit little bias. ¤

7.3 Categorical Forecasts of Continuous Predictands

A different set of verification measures is generally applied to forecasts of continuous atmospheric variables. Continuous variables can take on any value in a specified segment of the real line, rather than being limited to a finite number of discrete points. Temperature is an example of a continuous variable. In practice, however, forecasts and observations of continuous atmospheric variables are made using a finite number of discrete values. For example, temperature forecasts

are usually rounded to integer degrees. It would be possible to deal with this kind of forecast verification data in discrete form; however, there are usually so many allowable values of forecasts and observations that the resulting contingency tables would become quite unwieldy. Just as discretely reported observations of continuous atmospheric variables were treated as continuous quantities in Chapter 4, it is convenient and useful to treat the verification of (discrete) forecasts of continuous quantities in a continuous framework as well.

Conceptually, the joint distribution of forecasts and observations is again of fundamental interest. This distribution will be the continuous analog of the discrete joint distribution of Eq. (7.1). Because of the finite nature of the verification data, however, explicitly using the concept of the joint distribution in a continuous setting requires that a parametric distribution such as the bivariate normal [Eq. (4.42)] be assumed and fit. Parametric distributions are occasionally assumed for the joint distribution of forecasts and observations (e.g., Katz *et al.*, 1982), but it is far more common that scalar performance and skill measures, computed using individual forecast/observation pairs are used in verification of continuous categorical forecasts.

7.3.1 Conditional Quantile Plots

It is possible to graphically represent certain aspects of the joint distribution of categorical forecasts and observations for continuous variables. For example, Fig. 7.4 shows *conditional quantile plots* for a sample of daily maximum temperature forecasts issued during the winters of 1980/81 through 1985/86 for Minneapolis, Minnesota. Panel a represents the performance of objective (MOS) forecasts, and panel b represents the performance of the corresponding subjective forecasts. These diagrams contain two parts, representing the two factors in the calibration-refinement factorization of the joint distribution of forecasts and observations [Eq. (7.2)]. The conditional distributions of the observations given the forecasts are represented in terms of selected quantiles, in comparison to the 1 : 1 diagonal line representing perfect forecasts. Here it can be seen that the MOS forecasts (panel a) exhibit a small degree of overforecasting (the conditional medians of the observed temperatures are consistently colder than the forecasts), whereas the subjective forecasts are essentially unbiased. The histograms in the lower parts of the panels represent the frequency of use of the forecasts, or $p(y_i)$. Here it can be seen that the subjective forecasts are somewhat sharper, with more extreme temperatures being forecast more frequently, especially on the left tail.

7.3.2 Scalar Accuracy Measures

Only two scalar measures of forecast accuracy for continuous predictands are in common use. The first is the *mean absolute error* (MAE):

$$\mathrm{MAE} = \frac{1}{n} \sum_{k=1}^{n} |y_k - o_k|. \tag{7.14}$$

Fig. 7.4 Conditional quantile plots for objective (a) and subjective (b) 24-h categorical maximum temperature forecasts, for winter seasons of 1980–1986 at Minneapolis, Minnesota. Main bodies of the figures delineate smoothed conditional quantiles from the conditional distributions $p(o_j \mid y_i)$ in relation to the 1 : 1 line, and the lower parts of the figures show the unconditional distribution of the forecasts, $p(y_i)$. [From Murphy *et al.* (1989), reproduced with permission of the American Meteorological Society.]

Here (y_k, o_k) is the kth of n pairs of forecasts and observations, and the MAE is the arithmetic average of the absolute values of the differences between the members of each pair. Clearly, the MAE is zero if the forecasts are perfect (each $y_k = o_k$) and increases as discrepancies between the forecasts and observations become larger. One can interpret the MAE as a typical magnitude for the forecast error in a given verification data set.

Fig. 7.5 Year-by-year MAE for October–March objective maximum temperature forecasts at the 24- and 48-h projections, for approximately 90 stations in the United States. Forecasts for 1970/71–1972/73 were produced by perfect-prog equations; those for 1973/74 onward were produced by MOS equations. [From Carter *et al.* (1989), reproduced with permission of the American Meteorological Society.]

The MAE is often used as a verification measure for temperature forecasts in the United States. Figure 7.5 shows MAE for objective maximum temperature forecasts at approximately 90 stations in the United States during the cool seasons (October–March) 1970/71 through 1987/88. Temperature forecasts with a 24-h lead time are more accurate than those for a 48-h lead time, and exhibit smaller average absolute errors. A clear trend of forecast improvement through time is also evident, as the MAE for the 48-h forecasts in the 1980s is comparable to the MAE for the 24-h forecasts in the early 1970s.

The other common accuracy measure for continuous categorical forecasts is the *mean-squared error* (MSE):

$$\mathrm{MSE} = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2. \tag{7.15}$$

The MSE is the average squared difference between the forecast and observation pairs. This measure is similar to the MAE except that the squaring function is used rather than the absolute value function. Since the MSE is computed by squaring forecast errors, it will be more sensitive to larger errors than will the MAE. Squaring the errors necessarily produces positive terms in Eq. (7.15), so the MSE increases from zero for perfect forecasts through larger positive values as the dis-

crepancies between forecasts and observations become increasingly large. Sometimes the MSE is expressed as its square root, RMSE = $\sqrt{\text{MSE}}$, which has the same physical dimensions as the forecasts and observations, and can also be thought of as a typical magnitude for forecast errors.

Initially, one might think that the correlation coefficient [Eq. (3.17)] could be another useful accuracy measure for categorical forecasts of continuous predictands. The correlation does reflect linear association between two variables (in this case, forecasts and observations), but it does not account for biases that may be present in the forecasts. This can be appreciated by considering an algebraic manipulation of the MSE (Murphy, 1988):

$$\text{MSE} = (\bar{y} - \bar{o})^2 + s_y^2 + s_o^2 - 2s_y s_o r_{yo}. \tag{7.16}$$

Here r_{yo} is the product–moment correlation between the forecasts and observations, s_y and s_o are the standard deviations of the marginal distributions of the forecasts and observations, respectively, and the first term in Eq. (7.16) is the square of the mean error (ME)

$$\text{ME} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k) = \bar{y} - \bar{o}. \tag{7.17}$$

The mean error is simply the difference between the average forecast and average observation, and therefore expresses the bias of the forecasts. Equation (7.17) differs from Eq. (7.15) in that the individual forecast errors are not squared before they are averaged. Forecasts that are on average too high will exhibit ME > 0, and forecasts that are on average too low will exhibit ME < 0. It is important to note that the bias gives no information about the typical magnitude of individual forecast errors, and is therefore not an accuracy measure.

Returning to Eq. (7.16), it can be seen that forecasts that are more highly correlated with the observations will exhibit lower MSE, other factors being equal. However, since the MSE can be written with the correlation r_{yo} and the bias (ME) in separate terms, one can imagine forecasts that are highly correlated with the observations, but with sufficiently severe bias that they would be useless at face value. A set of temperature forecasts could exist, for example, that are exactly half of the subsequently observed temperature. For convenience, imagine that these temperatures are nonnegative. A scatterplot of the observed temperatures versus the corresponding forecasts would exhibit all points falling perfectly on a straight line ($r_{yo} = 1$), but the slope of that line would be 2. The bias, or mean error, would be ME $= n^{-1}\Sigma_k(f_k - o_k) = n^{-1}\Sigma_k(0.5o_k - o_k)$, or minus half of the average observation. This bias would be squared in Eq. (7.16), leading to a very large MSE. A similar situation would result if all the forecasts were exactly 10° colder than the observed temperatures. The correlation r_{yo} would still be one, the points on the scatterplot would fall on a straight line (this time with unit slope), the ME would be $-10°$, and the MSE would be inflated by $(10°)^2$.

7.3.3 Skill Scores

Relative accuracy measures of the form of Eq. (7.4) can easily be constructed using the MAE or MSE. Usually the reference, or control, forecasts are provided either by the climatological values of the predictand or by persistence. For the MSE, the accuracies of these two references are, respectively,

$$\mathrm{MSE}_{\mathrm{Clim}} = \frac{1}{n} \sum_{k=1}^{n} (\bar{o} - o_k)^2 \tag{7.18a}$$

and

$$\mathrm{MSE}_{\mathrm{Pers}} = \frac{1}{n} \sum_{k=1}^{n} (o_{k-1} - o_k)^2, \tag{7.18b}$$

where subscripts Clim and Pers indicate climatology and persistence, respectively. Completely analogous equations can be written for the MAE, in which the squaring function would be replaced by the absolute value function.

In Eq. (7.18a), it is implied that the climatological average value does not change from forecast occasion to forecast occasion (i.e., as a function of the index, k). If this is true, then $\mathrm{MSE}_{\mathrm{Clim}}$ in Eq. (7.18a) is an estimate of the sample variance of the predictand [cf. Eq. (3.5)]. In some applications the climatological value of the predictand will be different for different forecasts. For example, if daily temperature forecasts were being verified over the course of several months, the index k would represent time, and the climatological average temperature would usually change smoothly as a function of the date. In this case the quantity being summed in Eq. (7.18a) would be $(c_k - o_k)^2$, where c_k represents the climatological value of the predictand on day k. Similarly, the MSE for persistence (7.18b) implies that the index k represents time, so that the reference forecast for the observation o_k at time k is just the observation of the predictand during the previous time period, o_{k-1}.

Either of the reference measures of accuracy in Eq. (7.18a) or (7.18b), or their MAE counterparts can be used in Eq. (7.4). For the MSE using climatology as the control forecasts, the skill score (in proportion rather than percentage terms) becomes

$$\mathrm{SS}_{\mathrm{Clim}} = \frac{\mathrm{MSE} - \mathrm{MSE}_{\mathrm{Clim}}}{0 - \mathrm{MSE}_{\mathrm{Clim}}} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{Clim}}}. \tag{7.19}$$

Notice that perfect forecasts have MSE or MAE = 0, which allows use of the rearrangement of the skill score in Eq. (7.19).

Example 7.5. Skill of the Temperature Forecasts in Fig. 7.5

The analog of Eq. (7.19) for the MAE can be applied to the temperature forecast accuracy data in Fig. 7.5. Assume the reference MAE is $\mathrm{MAE}_{\mathrm{Clim}} = 8.5°\,\mathrm{F}$. This

value will not depend on the forecast projection, and should be different for different years only to the extent that the average MAE values plotted in the figure are for slightly different collections of stations.

For 1986/87 the MAE for the 24-h projection is 3.5° F, yielding a skill score of $SS_{Clim} = 1 - (3.5°\,F)/(8.5°\,F) = 0.59$, or a 59% improvement over climatological forecasts. For the 48-h projection the MAE is 4.3° F, yielding $SS_{Clim} = 1 - (4.3°\,F)/(8.5°\,F) = 0.49$, or a 49% improvement over climatology. Not surprisingly, the forecasts for the 24-h projection are more skillful than those for the 48-h projection. ¤

The skill score for the MSE in Eq. (7.19) can be manipulated algebraically in a way that yields some insight into the determinants of forecast skill, as measured by the MSE with respect to climatology as the reference [Eq. (7.18a)]. Rearranging Eq. (7.19), and substituting an expression for the Pearson product–moment correlation between the forecasts and observations, r_{yo}, we obtain (Murphy, 1988)

$$SS_{Clim} = r_{yo}^2 - \left[r_{yo} - \frac{s_y}{s_o} \right]^2 - \left[\frac{\bar{y} - \bar{o}}{s_o} \right]^2. \tag{7.20}$$

Equation (7.20) indicates that the skill in terms of the MSE can be regarded as consisting of a contribution due to the correlation between the forecasts and observations, and penalties relating to the reliability and bias of the forecasts.

The first term in Eq. (7.20) is the square of the product–moment correlation coefficient, and is a measure of the proportion of variability in the observations that is (linearly) accounted for by the forecasts (Chapter 6).

The second term in Eq. (7.20) is a measure of reliability, or conditional bias, of the forecasts. This is most easily appreciated by imagining a linear regression between the observations and the forecasts. The slope, b, of a linear regression equation can be expressed in terms of the correlation and the standard deviations of the predictor and predictand as $b = (s_o/s_y)r_{yo}$. This can be verified by substituting Eqs. (3.5) and (3.18) into Eq. (6.7a). If this slope is smaller than $b = 1$, the predictions made with this regression are too large (positively biased) for smaller forecasts and too small (negatively biased) for larger forecasts. However, if $b = 1$, there will be no conditional bias, and substituting $b = (s_o/s_y)r_{yo} = 1$ into the second term in Eq. (7.20) will yield a zero penalty for conditional bias.

The third term in Eq. (7.20) is the square of the unconditional bias, as a fraction of the standard deviation of the observations, s_o. If the bias is small compared to the variability of the observations as measured by s_o, the reduction in skill will be modest, whereas increasing bias of either sign progressively degrades the skill.

Thus, if the forecasts are completely reliable and unbiased, the second two terms in Eq. (7.20) are both zero, and the skill score is exactly r_{yo}^2. To the extent that the forecasts are biased or not completely reliable (in the sense of exhibiting conditional bias), then the square of the correlation coefficient will overestimate skill.

7.3.4 Skill in Probability Space

The skill of categorical forecasts for continuous predictands is almost always assessed on the basis of either the MAE or the MSE. A relatively new and interesting alternative is based on the *linear error in probability space* (LEPS) (Ward and Folland, 1991) score. As of this writing it is too soon to judge the extent to which LEPS-based skill scores will be useful or well accepted, but describing it here serves at minimum to illustrate that possibilities for verification of continuous categorical forecasts other than the MAE and MSE do exist.

The basic measure of the LEPS score is not linear or squared in the dimensions of the variable being forecast and observed, but rather is linear in the cumulative probability of that variable. That is, forecast errors are assessed with respect to their differences on the scale of the climatological cumulative distribution function, $F_o(o)$ rather than their magnitudes in the ordinary sense. The idea of measuring error in probability terms is appealing because forecasts of a quantity in regions of its probability distribution where it is relatively more likely should be easier than for values that are climatologically less likely. Transforming forecasts and observations to the cumulative probability scale assesses larger penalties for forecast errors in regions where the probability is more strongly concentrated because the steepness of the cumulative distribution function is greatest in these regions.

As before, let y_k and o_k be the kth forecast and observation pair. If $F_o(o)$ is the climatological cumulative distribution function for the observations, LEPS is the absolute forecast error in probability units, given by $|F_o(y_k) - F_o(o_k)|$. This is an accuracy measure that ranges from zero for perfect forecasts, and approaches one for the worst possible forecasts. Given a reference forecast, it is straightforward to construct a skill score of the form of Eq. (7.4). Here the simplest reference is the climatological median, for which $F_o(o_{\text{median}}) = 0.5$. Using Eq. (7.4), then, the LEPS skill score for a collection of n forecasts with respect to climatology is

$$\begin{aligned} \mathrm{SS}_{\mathrm{LEPS}} &= \frac{\Sigma_{k=1}^{n} |F_o(y_k) - F_o(o_k)| - \Sigma_{k=1}^{n} |0.5 - F_o(o_k)|}{0 - \Sigma_{k=1}^{n} |0.5 - F_o(o_k)|} \\ &= 1 - \frac{\Sigma_{k=1}^{n} |F_o(y_k) - F_o(o_k)|}{\Sigma_{k=1}^{n} |0.5 - F_o(o_k)|}. \end{aligned} \tag{7.21}$$

Figure 7.6 illustrates the concept of measuring accuracy in probability space with a hypothetical cumulative distribution function $F_o(o)$ for a variable that is strongly skewed to the right. The concentration of probability at small values is reflected by the steep initial slope of this cumulative distribution function. Solid arrows and dashed lines compare MAE and LEPS measures of forecast accuracy for forecasts on the low (y_k^-) and high (y_k^+) sides of an observation o_k. The absolute error for the y_k^- forecast is lower, but the fact that it is in a more probable region of the climatological distribution results in its accuracy being less accord-

Fig. 7.6 Hypothetical climatological cumulative distribution function $F_o(x)$ for a strongly positively skewed variable, O; and the relationships between MAE (horizontal arrows) and LEPS (vertical arrows) measures of errors for forecasts on the low (y_k^-) and high (y_k^+) sides of an observation o_k. The absolute error for the y_k^- forecast is smaller, but the fact that it is in a more probable region of the climatological distribution results in it being assessed lower accuracy in probability space. Also shown (lightface arrows) are MAE and LEPS measures of error for the climatological median forecast.

ing to the LEPS measure. Also shown (lighter arrows) are the MAE and LEPS measures of accuracy for the climatological median reference forecast.

7.4 Probability Forecasts

Verification of probability forecasts is somewhat more subtle than verification of categorical forecasts. Since categorical forecasts contain no expression of uncertainty, it is clear whether an individual forecast is correct. However, unless a probability forecast is either 0.0 or 1.0, the situation is less clear-cut. For probability values between these two (certainty) extremes a single forecast is neither "right" or "wrong," so that meaningful assessments can only be made on collections of forecast and observation pairs. Again, it is the joint distribution of forecasts and observations that contains the relevant information for forecast verification.

As a practical matter, probability forecasts are formulated and verified for discrete predictands. Although some predictands (such as precipitation vs. no pre-

cipitation) are naturally discrete, verification of probabilistic forecasts for physically continuous variables (such as temperature or precipitation amount) require construction of MECE categories of the predictand that can be assigned finite probability. Methods for verification of probability forecasts are divided according to whether the discrete predictand is dichotomous (binary), or is polychotomous (represented by three or more categories).

7.4.1 Dichotomous Events

Dichotomous events are limited to $J = 2$ possible outcomes. The most familiar example of probability forecasts for a dichotomous event is the probability of precipitation (PoP) forecast. Here the event is either the occurrence (o_1) or nonoccurrence (o_2) of measurable precipitation. The joint distribution of forecasts and observations is more complicated than for the case of categorical forecasts of binary predictands, however, because more than $I = 2$ probability values can allowably be forecast. In theory any real number between zero and one is an allowable probability forecast, but in practice the forecasts are usually rounded to one of a reasonably small number of values.

Table 7.3 contains a verification data set for $n = 12{,}402$ subjective PoP forecasts made for locations in the United States between October 1980 and March 1981. The forecasters were required to round their subjective probabilities for the occurrence of precipitation to one of $I = 12$ possible values, y_i. Thus, the joint distribution of forecasts and observations for this data set contains $I \times J = 24$ individual probabilities. Table 7.3 presents this joint distribution in terms of the calibration-refinement factorization [Eq. (7.2)]. That is, for each possible forecast probability, y_i, Table 7.3 shows the relative frequency with which that forecast value was used, $p(y_i)$, and the conditional probability that precipitation occurred given that forecast, $p(o_1 \mid y_i)$, $i = 1, \ldots, I$. Because the predictand is binary it is not necessary to specify the conditional probability of precipitation not occurring given each of the forecasts. That is, the two predictand values represented by o_1 and o_2 constitute a MECE partition of the sample space, so that $p(o_2 \mid y_i) = 1 - p(o_1 \mid y_i)$. Similarly, not all of the $J = 12$ probabilities in the marginal distribution of the forecasts can be specified independently, since $\Sigma_j p(y_j) = 1$. Thus the joint distribution can be completely specified with $I \times J - 1 = 23$ of the 24 probabilities given in Table 7.3, which is the dimensionality of this verification problem.

7.4.2 Brier Score

Given the generally high dimensionality of verification involving probability forecasts even for dichotomous predictands, it is not surprising that forecast performance is most often assessed with a scalar summary measure. While attractive from a practical standpoint, such simplifications necessarily will give incomplete pictures of forecast performance. A number of scalar accuracy measures for verification of probabilistic forecasts of dichotomous events exist (Murphy and Daan,

1985), but by far the most common is the Brier score (BS). The Brier score is essentially the mean-squared error of the probability forecasts, considering that the observation is $o_1 = 1$ if the event occurs and $o_2 = 0$ if the event does not occur. The score averages the squared differences between pairs of forecast probabilities and the subsequent binary observations,

$$\mathrm{BS} = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2, \tag{7.22}$$

where the index k again denotes a numbering of the n forecast/event pairs. Comparing the Brier score with Eq. (7.15) for the MSE, it can be seen that the two are completely analogous. As a MSE measure of accuracy, the Brier score is negatively oriented, with perfect forecasts exhibiting BS = 0. Less accurate forecasts receive higher Brier scores, but since individual forecasts and observations are both bounded by zero and one, the score can take on values only in the range $0 \leq \mathrm{BS} \leq 1$.

The Brier score as expressed in Eq. (7.22) is used nearly universally; however, it differs from the score as originally introduced by Brier (1950) in that it averages only the squared differences pertaining to one of the two binary events. The original Brier score also included squared differences for the complementary (or "non-") event in the average, with the result that Brier's original score is exactly twice that given by Eq. (7.22). The confusion is unfortunate, but the usual present-day understanding of the meaning of *Brier score* is that given in Eq. (7.22). In order to distinguish this from the original formulation, the Brier score in Eq. (7.22) is sometimes referred to as the *half-Brier score*.

Skill scores of the form of Eq. (7.4) are often computed for the Brier score,

$$\mathrm{SS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{0 - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{7.23}$$

since $\mathrm{BS}_{\mathrm{perf}} = 0$. Usually the reference forecasts are the relevant climatological relative frequencies. Skill scores with respect to the climatological probabilities for subjective PoP forecasts during the warm seasons of 1966–1986 are shown in Fig. 7.7. The labeling of the vertical axis as "% improvement over climate" indicates that it is the skill score in Eq. (7.23), using climatological probabilities as the reference forecasts, that is plotted in the figure. Evidently forecasts made at the 36-h projection in the mid-1980s exhibited skill equivalent to 24-h forecasts made in the mid-1960s. Similarly, forecasts made at the 48-h projection in the mid-1980s were somewhat more skillful than the 36-h forecasts in the mid-1960s.

7.4.3 Algebraic Decomposition of the Brier Score

An instructive algebraic decomposition of the Brier score [Eq. (7.22)] has been derived by Murphy (1973). It relates to the calibration-refinement factorization of

Fig. 7.7 Trends in the skill of U.S. subjective PoP forecasts, measured in terms of the Brier score with respect to climatological probabilities, April–September 1966–1985. Solid lines show least-squares regression lines fit to the skill scores. [From Carter and Polger (1986).]

the joint distribution $p(y_i, o_j)$, Eq. (7.2), in that it pertains to quantities that are conditional on particular values of the forecasts.

As before, consider that a verification data set contains forecasts taking on any of a discrete number, I, of forecast values y_i. For example, in the verification data set in Table 7.3, there are $I = 12$ allowable forecast values, ranging from $y_1 = 0.0$ to $y_{12} = 1.0$. Let N_i be the number of times each forecast y_i is used in the collection of forecasts being verified. The total number of forecast/event pairs is simply the sum of these subsample, or conditional sample, sizes:

$$n = \sum_{i=1}^{I} N_i. \tag{7.24}$$

The marginal distribution of the forecasts, or the "refinement" in the calibration-refinement factorization, consists simply of the relative frequencies

$$p(y_i) = \frac{N_i}{n}. \tag{7.25}$$

The third row in Table 7.3 shows these relative frequencies for the data set represented there.

For each subsample delineated by the I allowable forecast values there is a relative frequency of occurrence of the forecast event. Since the observed event is dichotomous, a single conditional relative frequency defines the conditional distribution of observations given each forecast y_i. It is convenient to think of this

relative frequency as the subsample relative frequency, or conditional average observation,

$$\bar{o}_i = p(o_1 \mid y_i) = \frac{1}{N_i} \sum_{k \in N_i} o_k, \tag{7.26}$$

where $o_k = 1$ if the event occurs for the kth forecast/event pair and $o_k = 0$ if it does not, and the summation is over only those values of k corresponding to occasions when the forecast y_i was issued. The second row in Table 7.3 shows these conditional relative frequencies. Similarly, the overall (unconditional) relative frequency, or sample climatology, of the observations is given by

$$\bar{o} = \frac{1}{n} \sum_{k=1}^{n} o_k. \tag{7.27}$$

After some algebra, the Brier score in Eq. (7.22) can be expressed in terms of the quantities defined above as the sum of the three terms

$$\mathrm{BS} = \underbrace{\frac{1}{n} \sum_{i=1}^{I} N_i (y_i - \bar{o}_i)^2}_{\text{(reliability)}} - \underbrace{\frac{1}{n} \sum_{i=1}^{I} N_i (\bar{o}_i - \bar{o})^2}_{\text{(resolution)}} + \underbrace{\bar{o}(1 - \bar{o})}_{\text{(uncertainty)}}. \tag{7.28}$$

As indicated in this equation, these terms are known as *reliability*, *resolution*, and *uncertainty*. Since more accurate forecasts are characterized by smaller values of BS, the forecaster would like the reliability term to be as small as possible, and the resolution term to be as large (in an absolute value sense) as possible. Equation (7.27) indicates that the uncertainty term depends only on the sample climatological relative frequency, and is unaffected by the forecasts.

The *reliability* term in Eq. (7.28) summarizes the calibration, or conditional bias, of the forecasts. It consists of a weighted average of the squared differences between the forecast probabilities y_i and the relative frequencies of the forecast event in each subsample. For forecasts that are perfectly reliable, the subsample relative frequency is exactly equal to the forecast probability in each subsample. The relative frequency of the forecast event should be small on occasions when $y_1 = 0.0$ is forecast, and should be large when $y_1 = 1.0$ is forecast. On those occasions when the forecast probability is 0.5, the relative frequency of the event should be near 1/2. For reliable, or well-calibrated forecasts, all the squared differences in the reliability term will be near zero, and their weighted average will be small.

The *resolution* term in Eq. (7.28) summarizes the ability of the forecasts to discern subsample forecast periods with different relative frequencies of the event. The forecast probabilities y_i do not appear explicitly in this term, yet it still depends on the forecasts through the sorting of the events making up the subsample relative frequencies [Eq. (7.26)]. Mathematically, the resolution term is a weighted average of the squared differences between these subsample relative frequencies, and the overall sample climatological relative frequency. Thus, if the forecasts sort

the observations into subsamples having substantially different relative frequencies than the overall sample climatology, the resolution term will be large. This is a desirable situation, since the resolution term is subtracted in Eq. (7.28). Conversely, if the forecasts sort the events into subsamples with very similar event relative frequencies, the squared differences in the summation of the resolution term will be small. In this case the forecasts resolve the event only weakly, and the resolution term will be small.

The *uncertainty* term in Eq. (7.28) depends only on the variability of the observations, and cannot be influenced by anything the forecaster may do. This term is identical to the variance of the binomial distribution (Chapter 4), exhibiting minima at zero when the climatological probability is either zero or one, and a maximum when the climatological probability is 0.5. When the event being forecast almost never happens, or almost always happens, the uncertainty in the forecasting situation is small. In this case, always forecasting the climatological probability will give generally good results. When the climatological probability is close to 0.5, there is substantially more uncertainty inherent in the forecasting situation, and the third term in Eq. (7.28) is commensurately larger.

7.4.4 The Attributes Diagram

The terms in Eq. (7.28) can be interpreted geometrically with respect to a device known as the *attributes diagram* (Hsu and Murphy, 1986). An example of an attributes diagram is shown in Fig. 7.8. The plotted points in the attributes diagram show the subsample relative frequencies [Eq. (7.26)] as a function of the forecasts y_i. For perfectly reliable forecasts these paired quantities are equal, yielding all points in the diagram falling on the diagonal line labeled "Perfect reliability." Ordinarily the plotted points are connected by line segments to make the diagram easier to read and to contrast with the diagonal 1 : 1 line.

The reliability term in Eq. (7.28) can be seen to consist of the weighted average of the squared vertical distances between the points and the diagonal line. The sample of PoP forecasts plotted in Fig. 7.8 shows good reliability, with the plotted points falling quite close to the 1 : 1 line. It is not unusual to find such good reliability for subjective weather forecasts in cases where the forecasters are experienced at probability assessment for the weather element being forecast.

The resolution term can also be interpreted geometrically with reference to the attributes diagram in Fig. 7.8. Geometrically, the identification of subsets with event relative frequencies different from the climatological probability corresponds to points in the attributes diagram being well removed, vertically, from the level of the overall sample climatology. This level is indicated in the figure by the "No resolution" line. Forecasts for which the points fall on the "No resolution" line are unable to resolve occasions where the event is more or less likely than the overall climatological probability. The weighted average constituting the resolution term is of the squares of the vertical distances between the points (the sub-

Fig. 7.8 Attributes diagram for the PoP forecasts summarized in Table 7.3. Solid dots show observed relative frequency of precipitation occurrence, conditional on each of the $I = 12$ possible probability forecasts. Perfectly calibrated forecasts would exhibit all points falling on the diagonal line labeled "Perfect reliability." Forecasts not defining event subsets with different relative frequencies of the forecast event would exhibit all points on the dashed "No resolution" line, which is plotted at the level of the sample climatological probability. Points in the stippled region bounded by the dotted line labeled "No skill" contribute positively to forecast skill. Relative frequency of use of each of the forecast values, $p(y_i)$, are shown parenthetically.

sample relative frequencies) and the "No resolution" line. These distances will be large for forecasts exhibiting good resolution, in which case the resolution term will contribute to a small (i.e., good) Brier score. The forecasts summarized in Fig. 7.8 exhibit a substantial degree of resolution, with forecasts that are most different from the sample climatological probability of 0.162 making the largest contributions to the resolution term.

Another interpretation of the uncertainty term emerges from imagining the attributes diagram for climatological forecasts. Since only a single forecast value (the climatological probability) is ever used in this case, there is only $I = 1$ dot on the diagram. The horizontal position of this dot is at the climatological probability

for the event, and if the sample size is sufficiently large, the long-term climatological probability will be close to the sample climatological relative frequency, Eq. (7.27). The vertical position of the single dot will be at the sample climatological relative frequency, locating it at the intersection of the "Perfect reliability," "No skill," and "No resolution" lines. Thus, climatological forecasts have perfect [zero, in Eq. (7.28)] reliability, since the "forecast" and the conditional relative frequency [Eq. (7.26)] are both equal to the climatological probability [Eq. (7.27)]. Similarly, the climatological forecasts have zero resolution since the existence of only $I = 1$ forecast category precludes discerning different subsets of forecasting occasions with differing relative frequencies of the outcomes. Since the *reliability* and *resolution* terms in Eq. (7.28) are both zero, it is clear that the Brier score for climatological forecasts is exactly the *uncertainty*.

This observation of the equivalence of the uncertainty term and the BS for climatological forecasts has interesting consequences for the skill score in Eq. (7.23). Substituting Eq. (7.28) for BS, and *uncertainty* for $\mathrm{BS}_{\mathrm{Clim}}$ yields

$$\mathrm{SS} = \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}}. \tag{7.29}$$

Since the uncertainty term is always positive, the probability forecasts will exhibit positive skill in the sense of Eq. (7.23) if the resolution term is larger in absolute value than the reliability term. This means that subsamples of the forecasts will contribute positively to the overall skill when their resolution term is larger than their reliability term. Geometrically, this corresponds to points on the attributes diagram being closer to the 1 : 1 "Perfect reliability" line than to the horizontal "No resolution" line. This condition defines the "No skill" line, which is midway between the "Perfect reliability" and "No resolution" lines, and delimits the stippled region, in which subsamples contribute positively to forecast skill. In Fig. 7.8 only the subsample for $y_4 = 0.2$, which is nearly the climatological probability, fails to contribute positively to the overall forecast skill.

7.4.5 The Reliability Diagram

Often a simpler form of the attributes diagram, called the *reliability diagram*, is used to graphically represent the performance of probability forecasts of dichotomous events. A reliability diagram consists of only the plot of observed relative frequency as a function of forecast probability, the 1 : 1 diagonal perfect reliability line, and a summary of the frequency of use of each forecast value. The reliability diagram is less cluttered than an attributes diagram, allowing a more prominent display of the frequency of use of the forecasts. This is an important consideration, since the plotted points in the reliability or attributes diagrams represent the conditional distribution of observations given the forecasts, $p(o_j \mid y_i)$, and the frequency of use of the forecasts is just the unconditional distribution of the forecasts, $p(y_i)$. Thus the reliability diagram is a compact display of the full distribu-

Fig. 7.9 Hypothetical reliability diagrams showing observed relative frequency as a function of forecast probabilities for (a) climatological forecasts, (b) forecasts exhibiting minimal resolution, (c) forecasts showing an underforecasting bias, (d) forecasts showing good resolution at the expense of reliability, (e) reliable forecasts of a rare event, and (f) a verification data set limited by small sample size. Inset boxes indicate frequency of use of the forecasts, $p(y_i)$, and complete the graphical representation of the calibration-refinement factorization of the joint distribution of forecasts and observations.

tion of forecasts and observations through the calibration-refinement factorization [Equation (7.2)], and as such is a much more informative representation of forecast performance than are single scalar scores.

Figure 7.9 shows six hypothetical reliability diagrams, illustrating the ease with which a variety of kinds of relationships between forecasts and observations can be diagnosed. In each case, the inset rectangle contains a histogram portraying the relative frequency of use of the forecasts. Figure 7.9a shows the simple case of climatological forecasts. Here there is only one forecast, as indicated by the single dot, and the single histogram bar. Of course, it shows perfect reliability, but no resolution or skill.

The remaining panels in Fig. 7.9 show the use of the $I = 11$ equally spaced forecast probabilities 0.0, 0.1, 0.2, . . . , 1.0. The forecasts in Fig. 7.9b are evidently quite poor, showing minimal resolution. Those in Fig. 7.9c exhibit the same frequency of use of the forecasts, but with much better results. Here there is good resolution, but a consistent underforecasting bias. The forecasts in Fig. 7.9d also show excellent resolution, particularly considering the U-shaped distribution of $p(y_i)$. However, these sharper forecasts have come at the price of degraded reliability, with evident overforecasting of the smaller probabilities and underfore-

casting of the larger probabilities. Forecasters producing either Fig. 7.9c or 7.9d, seeing their work summarized in this way, could take definite steps to improve the quality of future forecasts. The real forecasts summarized in Fig. 7.8 show excellent reliability, but could be improved by increasing the resolution through (reliably) forecasting the higher probability values more frequently.

Figures 7.9e and 7.9f show two other common reliability diagram situations. Figure 7.9e is typical of well-calibrated forecasts of a relatively rare event. The inset histogram indicates that the lowest probability values are used most often, and that the higher probabilities are essentially never issued. The forecasts are highly reliable over their range of use, and further skill could be derived only from more frequent (but still reliable) use of higher probability values. Finally the reliability diagram in Fig. 7.9f is typical of verification data sets with insufficient sample size. The highly erratic behavior of the conditional relative frequencies indicates that the plot is dominated by sampling variations, and is of limited usefulness in assessing the quality of the forecasts.

7.4.6 Hedging and Strictly Proper Scoring Rules

When forecasts are quantitatively evaluated, it is natural for forecasters to want to achieve the best scores they can. Depending on the evaluation measure, it may be possible to improve scores by *hedging*, or forecasting something other than one's true beliefs about future weather events. In the setting of a "forecast contest" in a college or university, if the evaluation of one's performance can be improved by "playing the score," then it is entirely rational to try to do so. Conversely, if one is responsible for assuring that forecasts are of the highest possible quality, evaluating those forecasts in a way that penalizes hedging is desirable.

A forecast evaluation procedure that awards forecasters' best expected scores only when their true beliefs are forecast is called *strictly proper*. That is, strictly proper scoring procedures cannot be hedged. One very appealing attribute of the Brier score is that it is strictly proper, and this is one strong motivation for using the Brier score to evaluate the accuracy of probability forecasts. Of course, it is not possible to know in advance what Brier score a given forecast will achieve, unless one can make perfect forecasts. However, it is possible on each forecasting occasion to calculate the expected, or probability-weighted, score using one's subjective probability for the forecast event.

Suppose that a forecaster's subjective probability for the event being forecast is y^*, and that the forecaster must publicly issue a forecast probability, y. The expected Brier score is simply

$$E[\mathrm{BS}] = y^*(y - 1)^2 + (1 - y^*)(y - 0)^2, \tag{7.30}$$

where the first term is the score received if the event occurs multiplied by the subjective probability that it will occur, and the second term is the score received if the event does not occur multiplied by the subjective probability that it will not

occur. Consider that the forecaster has decided on a subjective probability y^*, and is weighing the problem of what forecast y to issue publicly. Regarding y^* as constant, it is easy to minimize the expected Brier score by differentiating Eq. (7.30) by y, and setting the result equal to zero. Then

$$\frac{\partial E[\mathrm{BS}]}{\partial y} = 2y^*(y - 1) + 2(1 - y^*)y = 0, \tag{7.31}$$

yielding

$$\begin{aligned} 2yy^* - 2y^* + 2y - 2yy^* &= 0, \\ 2y &= 2y^*, \end{aligned}$$

and

$$y = y^*.$$

That is, regardless of the forecaster's subjective probability, the minimum expected Brier score is achieved only when the publicly issued forecast corresponds exactly to the subjective probability. By contrast, the absolute error (linear) score, $\mathrm{LS} = |y_i - o_j|$ is minimized by forecasting $y = 0$ when $y^* < 0.5$, and forecasting $y = 1$ when $y^* > 0.5$.

7.4.7 Probability Forecasts for Multicategory Events

It is not necessary for the discrete events described by probability forecasts to be dichotomous, or "yes" versus "no." For some predictands forecast probability is distributed among three or more MECE events. These events may be *nominal*, for which there is not a natural ordering; or *ordinal*, when it is clear which events are larger or smaller than others. The approaches to verification of probability forecasts for nominal and ordinal predictands differ. This is because the magnitude of the forecast error is not a meaningful quantity in the case of nominal events, but is potentially quite important for ordinal events.

An example of probabilistic forecasts for nominal events is the conditional forecasts for freezing, frozen, or liquid precipitation in the FOUS14 MOS forecasts in Table 6.7. These give probabilities, conditional on occurrence of precipitation, that the precipitation will fall in one of three forms. The probabilities for freezing and frozen precipitation are given in the lines labeled POZP and POSN, respectively. Since mixed precipitation is operationally categorized as "frozen," the third event is all liquid precipitation, and these three events constitute a MECE partition of the (conditional) sample space. The conditional probability of liquid precipitation is 1 − POZP − POSN. Since there is no natural ordering by magnitudes of the three precipitation types, these are nominal predictands. The usual approach to verifying forecasts for nominal predictands is to collapse them to a sequence of binary predictands; in this case frozen versus not frozen, freezing versus not freezing, and liquid versus not liquid precipitation. Then, Brier scores,

attributes diagrams, and so on can be used to evaluate each derived binary forecasting situation.

Verification of probability forecasts for ordinal predictands presents a more difficult problem. First, the dimensionality of the verification problem increases exponentially with the number of categories over which the forecast probability is distributed. For example, consider a $J = 3$-event situation for which the forecast probabilities are constrained to be one of the 11 values 0.0, 0.1, 0.2, . . . , 1.0. The dimensionality of the problem is not simply $33 - 1 = 32$, as might be expected by extension of the dimensionality for the dichotomous forecast problem, because the forecasts are now vector quantities. For example, the forecast vector (0.2, 0.3, 0.5) is a different and distinct forecast from the vector (0.3, 0.2, 0.5). Since the three forecast probabilities must sum to 1.0, only two of them can vary freely. In this situation there are $I = 66$ possible forecast vectors, and a dimensionality for the forecast problem of $(66 \times 3) - 1 = 197$. Similarly, the dimensionality for the four-category ordinal verification situation with the same restriction on the forecast probabilities would be $(286 \times 4) - 1 = 1143$. As a practical matter, probability forecasts for ordinal predictands have been evaluated using only scalar performance measures, even though such approaches will necessarily be incomplete.

For ordinal predictands, collapsing the verification problem to a series of 2×2 tables will result in the loss of potentially important information related to the ordering of the outcomes. For example, the 30- and 90-day temperature forecasts discussed in Chapter 6 distribute probability among three MECE outcomes: cold, near-normal, and warm. If one were to verify the cold events in distinction to "not cold" events composed of both the near-normal and warm categories, information pertaining to the magnitude of the forecast error would be thrown away. That is, the same error magnitude would be assigned to the difference between cold and warm as to the difference between cold and near-normal.

7.4.8 Ranked Probability Score

What is desired for the verification of probability forecasts of ordinal predictands is a measure that is *sensitive to distance*; that is, the verification measure should be capable of penalizing forecasts increasingly as more probability is assigned to event categories further removed from the actual outcome. In addition, one would like the verification measure to be strictly proper, so that forecasters are encouraged to report their true beliefs. The most commonly used such measure is the *ranked probability score* (RPS) (Epstein, 1969b; Murphy, 1971). Many strictly proper scalar scores that are sensitive to distance exist (Murphy and Daan, 1985; Staël von Holstein and Murphy, 1978), but of these the RPS is preferred (Daan, 1985).

The RPS is essentially an extension of the Brier score to the multi-event situation. That is, it is a squared-error score with respect to the observation "1" if the

forecast event occurs, and "0" if the event does not occur. However, in order for the score to be sensitive to distance, the squared errors are computed with respect to the *cumulative* probabilities in the forecast and observation vectors. This characteristic introduces some notational complications.

As before, let J be the number of forecast categories, and therefore also the number of probabilities constituting each forecast. For example, the 30- and 90-day temperature forecasts discussed in Chapter 6 have $J = 3$ events over which to distribute probability. If the forecast is 20% chance of cold, 40% chance of near-normal, and 40% chance of warm, then $y_1 = 0.2$, $y_2 = 0.4$, and $y_3 = 0.4$. Similarly, the observation vector has three components. One of these components, corresponding to the event that occurs, will equal 1, and the other $J - 1$ components will equal zero. In the case of the long-range temperature forecasts, if the observed average temperature is in the warm category, then $o_1 = 0$, $o_2 = 0$, and $o_3 = 1$.

The cumulative forecasts and observations, denoted Y_m and O_m, are then defined as

$$Y_m = \sum_{j=1}^{m} y_j, \qquad m = 1, \ldots, J, \tag{7.32a}$$

and

$$O_m = \sum_{j=1}^{m} o_j, \qquad m = 1, \ldots, J. \tag{7.32b}$$

In terms of the example in the previous paragraph, $Y_1 = y_1 = 0.2$, $Y_2 = y_1 + y_2 = 0.6$, and $Y_3 = y_1 + y_2 + y_3 = 1.0$; and $O_1 = o_1 = 0$, $O_2 = o_1 + o_2 = 0$, and $O_3 = o_1 + o_2 + o_3 = 1$. Notice that since Y_m and O_m are both cumulative functions of probability components that must add to one, the last sums Y_J and O_J are always both equal to one by definition.

The RPS is the sum of squared differences between the components of the cumulative forecast and observation vectors in Eqs. (7.32a) and (7.32b), given by

$$\text{RPS} = \sum_{m=1}^{J} (Y_m - O_m)^2, \tag{7.33a}$$

or, in terms of the forecast and observed vector components f_j and o_j,

$$\text{RPS} = \sum_{m=1}^{J} \left[\left(\sum_{j=1}^{m} y_j \right) - \left(\sum_{j=1}^{m} o_j \right) \right]^2. \tag{7.33b}$$

A perfect forecast would assign all the probability to the single y_j corresponding to the event that subsequently occurs, so that the forecast and observation vectors would be the same. In this case, RPS = 0. Forecasts that are less than perfect receive scores that are positive numbers, so the RPS has a negative orientation.

Notice also that the final ($m = J$) term in Eqs. (7.33) is always zero, because the accumulations in Eqs. (7.32) ensure that $Y_J = O_J = 1$. Therefore, the worst possible score is $J - 1$. For $J = 2$, the RPS reduces to the Brier score [Eq. (7.22)]. Note that since the last term, for $m = J$, is always zero, as a practical matter it need not be computed.

Example 7.6. Mechanics of the Ranked Probability Score

Table 7.4 demonstrates the mechanics of computing the RPS, and illustrates the property of sensitivity to distance, for two hypothetical probability forecasts for precipitation amounts. Here the continuum of precipitation has been divided into $J = 3$ categories, <0.01 in., 0.01–0.24 in., and ≥0.25 in. Forecaster 1 has assigned the probabilities (0.2, 0.5, 0.3) to the three events, and forecaster 2 has assigned the probabilities (0.2, 0.3, 0.5). The two forecasts are similar, except that forecaster 2 has allocated more probability to the ≥0.25-in. category at the expense of the middle category. If no precipitation falls on this occasion, the observation vector will be that indicated in the table. Forecaster 1 should receive a better score in this case, since this forecaster assigned more probability closer to the observed category than did forecaster 2. The score for forecaster 1 is RPS = $(0.2-1)^2 + (0.7-1)^2 = 0.73$, and for forecaster 2 it is RPS = $(0.2-1)^2 + (0.5-1)^2 = 0.89$. The lower RPS for forecaster 1 indicates a better forecast.

If, on the other hand, some amount of precipitation greater than 0.25 in. had fallen, forecaster 2's probabilities would have been closer, and would have received the better score. The score for forecaster 1 would have been RPS = $(0.2-0)^2 + (0.7-0)^2 = 0.53$, and the score for forecaster 2 would have been RPS = $(0.2-0)^2 + (0.5-0)^2 = 0.29$. Note that in both of these examples, only the first two terms in Eq. (7.33) were needed to compute the RPS. ¤

Equation (7.33) gives the ranked probability score for only a single forecast/event pair. Jointly evaluating a collection of n forecasts using the RPS requires

Table 7.4

Comparison of Two Hypothetical Probability Forecasts for Precipitation Amount Divided into $J = 3$ Categories[a]

Event (inches)	Forecaster 1		Forecaster 2		Observed	
	y_j	Y_m	y_j	Y_m	o_j	O_m
<0.01	0.2	0.2	0.2	0.2	1	1
0.01 – 0.24	0.5	0.7	0.3	0.5	0	1
≥0.25	0.3	1.0	0.5	1.0	0	1

[a] The three components of the observation vector indicate that the observed precipitation was in the smallest category.

nothing more than averaging the RPS values for each forecast/event pair,

$$\overline{\mathrm{RPS}} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{RPS}_k. \tag{7.34}$$

Similarly, the skill score for a collection of RPS values relative to the RPS computed from the climatological probabilities can be computed as

$$S = \frac{\overline{\mathrm{RPS}} - \overline{\mathrm{RPS}}_{\mathrm{Clim}}}{0 - \overline{\mathrm{RPS}}_{\mathrm{Clim}}} = 1 - \frac{\overline{\mathrm{RPS}}}{\overline{\mathrm{RPS}}_{\mathrm{Clim}}}. \tag{7.35}$$

7.5 Categorical Forecasts of Fields

An important problem in forecast verification is characterization of the quality of forecasts of *atmospheric fields*, or spatial arrays of atmospheric variables. Forecasts for such fields as surface pressures, geopotential heights, and temperatures are produced routinely by weather forecasting centers worldwide. These forecasts are categorical, as probabilities or expressions of uncertainties are not included as part of the forecast format. An example of such an operational forecast is shown in Fig. 7.10a, which displays 24-h forecasts of sea-level pressures and 1000–500 mb thicknesses over a portion of the Northern Hemisphere, made on the morning of May 4, 1993 by the U.S. National Meteorological Center. Figure 7.10b shows the same fields as analyzed 24 h later. A subjective visual assessment of the two pairs of fields indicates that the main features correspond quite well, but that small discrepancies exist in the locations and magnitudes of the features.

Objective, numerical methods of verification for forecasts of atmospheric fields allow more quantitative assessments of forecast quality to be made. In practice, all such methods operate on *gridded* fields, or collections of values of the field variable sampled at, or interpolated to, a grid in the spatial domain. Usually the geographic grid consists of regularly spaced points either in distance, or in latitude and longitude.

Figure 7.11 illustrates the gridding process for a hypothetical pair of forecast and observed fields in a small spatial domain. Each field is represented in map form as contours, or isolines, of the mapped quantity. The grid imposed on each map is a regular array of points at which the fields are sampled. Here the grid consists of four rows in the north–south direction and five columns in the east–west direction. Thus the gridded forecast field consists of the $M = 20$ discrete values y_m, which represent the smoothly varying continuous forecast field. The gridded observed field consists of the $M = 20$ discrete values o_m, which represent the smoothly varying observed field at these same locations.

The accuracy of a field forecast is assessed by computing measures of the correspondence between the values y_m and o_m. If a forecast is perfect, then $y_m = o_m$ for each of the M gridpoints. Of course, gridded forecast and observed fields can

Fig. 7.10 Forecast (a) and subsequently analyzed (b) sea-level pressures (solid) and 1000–500-mb thicknesses (dashed) over a portion of the Northern Hemisphere for May 5, 1993.

Fig. 7.11 Hypothetical forecast (a) and observed (b) atmospheric fields represented as contour maps over a small rectangular domain. Objective assessment of the accuracy of the forecast requires that both the forecast and observed fields be "gridded," or interpolated to the same geographic grid (small circles). Here the grid has four rows in the north–south direction, and five columns in the east–west direction, so the forecast and observed fields are represented by the $M = 20$ discrete values y_m and o_m, respectively.

differ in many ways, even when there are only a small number of gridpoints. Put another way, the verification of field forecasts is a problem of very high dimensionality, even for small grids. Although examination of the joint distribution of forecasts and observation is in theory the preferred approach to verification of field forecasts, their large dimensionality suggests that this ideal may not be practically realizable. Rather, the correspondence between forecast and observed fields has been characterized using scalar summary measures. These scalar accuracy measures are necessarily incomplete, but are useful in practice.

7.5.1 S1 Score

The S1 score is an accuracy measure that is primarily of historical interest. It was designed to reflect the accuracy of the forecasts of gradients of pressure or geopotential height, in consideration of the relationship of these gradients to the wind field at the same level.

Rather than operating on individual gridded values, the S1 score operates on the differences between gridded values at adjacent gridpoints. Denote the differences between the gridded values at any particular pair of adjacent gridpoints as Δy for points in the forecast field, and Δo for points in the observed field. In terms of Fig. 7.11, for example, one possible value of Δy is $y_3 - y_2$, which would be compared to the corresponding gradient in the observed field, $\Delta o = o_3 - o_2$. Similarly, the difference $\Delta y = y_9 - y_4$ would be compared to the observed difference $\Delta o = o_9 - o_4$. If the forecast field reproduces the signs and magnitudes of the gradients in observed field exactly, each Δy will equal its corresponding Δo.

The S1 score summarizes the differences between the (Δy, Δo) pairs according to

$$S1 = \frac{\sum_{\substack{\text{adjacent}\\ \text{pairs}}} |\Delta y - \Delta o|}{\sum_{\substack{\text{adjacent}\\ \text{pairs}}} \max\{|\Delta y|, |\Delta o|\}} \times 100. \tag{7.36}$$

Here the numerator consists of the sum of the absolute errors in forecast gradient over all adjacent pairs of gridpoints. The denominator consists of the sum, over the same pairs of points, of the larger of the absolute value of the forecast gradient, $|\Delta y|$, or the absolute value of the observed gradient, $|\Delta o|$. The resulting ratio is multiplied by 100 only as a convenience.

Clearly, perfect forecasts will exhibit S1 = 0, and poorer forecasts will be characterized by increasingly larger scores. The S1 score has lost some favor because of certain undesirable characteristics. The most obvious is that the actual magnitudes of the forecast pressures or heights are unimportant, since only pairwise gridpoint differences are scored. Thus the S1 score does not reflect bias. Summer scores tend to be larger (apparently worse) because of weaker pressure gradients, producing a smaller denominator in Eq. (7.36). Finally, the score depends on the size of the domain and the spacing of the grid, so that it is difficult to compare S1 scores not pertaining to the same domain with the same grid.

Equation (7.36) yields the S1 score for a single pair of forecast/observed fields. When the aggregate skill of a series of field forecasts is to be assessed, the S1 scores for each forecast occasion are simply averaged. This averaging smooths sampling variations, and allows trends through time of forecast performance to be assessed more easily. For example, Fig. 7.12 shows average S1 scores for 36-h hemispheric 500-mb height forecasts, for January and July 1957–1988. The steady decline through time indicates improved forecast performance.

7.5.2 Mean-Squared Error

A much more common accuracy measure for field forecasts is the MSE, which operates on the gridded forecast and observed fields by simply averaging the individual squared differences between the two at each of the M gridpoints:

$$\text{MSE} = \frac{1}{M} \sum_{m=1}^{M} (y_m - o_m)^2. \tag{7.37}$$

This is essentially the same as Eq. (7.15), with the mechanics of both equations centered on averaging squared errors. The difference in application between the two equations is that the MSE in Eq. (7.37) is computed over the gridpoints of a *single* pair of forecast/observation fields—specifically, to $n = 1$ pair of maps—whereas Eq. (7.15) pertains to the average over n pairs of scalar forecasts and

Fig. 7.12 Average S1 scores for 36-h hemispheric forecasts of 500-mb heights for January and July, 1957–1988, produced by the Canadian Meteorological Centre. [From Stanski *et al.* (1989).]

observations. Clearly, the MSE for a perfectly forecast field is zero, with larger MSE indicating decreasing accuracy of the forecast.

Often the MSE is expressed as its square root, the root-mean squared error, RMSE = $\sqrt{\text{MSE}}$. This form of expression has the advantage that it retains the units of the forecast variable and is thus more easily interpretable as a typical error magnitude. To illustrate, the solid line in Fig. 7.13 shows RMSE in meters for 30-day forecasts of 500-mb heights initialized on 108 consecutive days during 1986/87). There is considerable variation in forecast accuracy from day to day, with the most accurate forecast fields exhibiting RMSE near 45 m, and the least accurate forecast fields exhibiting RMSE around 90 m. Also shown in Fig. 7.13 are RMSE values of 30-day forecasts of *persistence*, obtained by averaging observed 500-mb heights for the most recent 30 days prior to the forecast. Usually the persistence forecast exhibits slightly higher RMSE than the 30-day dynamical forecasts, but it is apparent from the figure that there are many days when the reverse is true, and that at this extended range the accuracy of these persistence forecasts was competitive with that of the dynamical forecasts.

The plot in Fig. 7.13 shows accuracy of individual field forecasts, but it is also possible to express the aggregate accuracy of a collection of field forecasts by averaging the MSEs for each of a collection of paired comparisons. This average of MSE values across many forecast maps can then be converted to an average RMSE as before, or expressed as a skill score in the same form as Eq. (7.19).

Fig. 7.13 Root-mean-square error (RMSE) for dynamical 30-day forecasts of 500-mb heights for the Northern Hemisphere between 20° and 80° N (solid), and persistence of the previous 30-day average 500-mb field (dashed), for forecasts initialized December 14, 1986–March 31, 1987. [From Tracton *et al.* (1989), reproduced with permission of the American Meteorological Society.]

Since the MSE for perfect field forecasts is zero, the skill score following the form of Eq. (7.4) is computed using

$$\mathrm{SS} = \frac{\Sigma_{k=1}^{n}\,\mathrm{MSE}(k) - \Sigma_{k=1}^{n}\,\mathrm{MSE}_{\mathrm{ref}}(k)}{0 - \Sigma_{k=1}^{n}\,\mathrm{MSE}_{\mathrm{ref}}(k)} = 1 - \frac{\Sigma_{k=1}^{n}\,\mathrm{MSE}(k)}{\Sigma_{k=1}^{n}\,\mathrm{MSE}_{\mathrm{ref}}(k)}, \tag{7.38}$$

where the aggregate skill of n individual field forecasts is being summarized. When this skill score is computed, the reference field forecast is usually either the climatological average field, or individual persistence forecasts as shown in Fig. 7.13.

7.5.3 Anomaly Correlation

The *anomaly correlation* (AC) is another commonly used measure of association that operates on pairs of gridpoint values in the forecast and observed fields. To compute the anomaly correlation, the forecast and observed values are first converted to *anomalies*, in the sense that the climatological average value of the observed field at each gridpoint is subtracted. The AC over the M gridpoints is computed according to

$$\mathrm{AC} = \frac{\Sigma_{m=1}^{M}\,[(y_m - C_m)(o_m - C_m)]}{[\Sigma_{m=1}^{M}\,(y_m - C_m)^2\,\Sigma_{m=1}^{M}\,(o_m - C_m)^2]^{1/2}}, \tag{7.39}$$

where C_m is the climatological value of the observed variable o_m at the point m, given by the average over n prior observations of the field

$$C_m = \frac{1}{n} \Sigma_{k=1}^{n} o_m(k). \tag{7.40}$$

The anomaly correlation is designed to detect similarities in the patterns of departures (i.e., anomalies) from the climatological mean field, and is sometimes referred to as a *pattern correlation.* If the forecast and observed fields are both above the climatological value at a particular gridpoint, the term for that gridpoint in the sum in the numerator of Eq. (7.39) will be positive. Similarly, if both forecast and observed values are below the climatological value, the term in the numerator for that gridpoint will be the product of two negative numbers, and will also contribute positively to the AC. To the extent that forecasts and observations occur on different sides of the climatological values, negative terms are introduced into the numerator, and the AC becomes smaller.

Superficially, the AC in Eq. (7.39) resembles the Pearson product–moment correlation coefficient [Eq. (3.17)]. The AC and the Pearson correlation do share the properties that both are bounded by ± 1, and that neither is sensitive to bias in the forecasts. However, there is an important difference between the AC and the Pearson correlation: The gridpoint-by-gridpoint climatological average values in Eq. (7.39) play the roles of the (constant) sample means in the Pearson correlation. The climatological values C_m in the AC are different at different gridpoints, and do not depend on the particular values of either the gridded forecasts or observations in the particular two fields being compared. By contrast, the sample means in the Pearson correlation are averages of the two samples being correlated, and are constant values. Although it is tempting to interpret the AC as the correlation between the forecast and observed anomalies, this interpretation is valid only if the averages over the M gridpoints of the two anomalies are zero, that is, only if $\Sigma_m(y_m - C_m) = 0$ and $\Sigma_m(o_m - C_m) = 0$. These conditions may be approximately valid if the forecast and observed fields are being compared over a large (e.g., hemispheric) domain, but will almost certainly not hold if the fields are compared over a relatively small area.

The AC is often used to evaluate extended-range (beyond a few days) forecasts. It conveys information that is similar to, although not exactly the same as, the MSE or RMSE. In particular, the MSE is more sensitive to the absolute magnitudes of the errors in the forecast fields. The AC is designed to reward good forecasts of the pattern of the observed field, with less sensitivity to the correct magnitudes of the field variable. Figure 7.14 shows AC values for the same 30-day dynamical and persistence forecasts of 500-mb height that are verified in terms of the RMSE in Fig. 7.13. Since the AC has a positive orientation (larger values indicate more accurate forecasts) and the RMSE has a negative orientation (smaller values indicate more accurate forecasts), one must mentally "flip" one of these two plots vertically in order to compare them. When this is done, it can

Fig. 7.14 Anomaly correlations for dynamical 30-day forecasts of 500-mb heights for the Northern Hemisphere between 20° and 80° N (solid), and persistence of the previous 30-day average 500-mb field (dashed), for forecasts initialized December 14, 1986–March 31, 1987. The same forecasts are evaluated in Fig. 7.13 using the RMSE. [From Tracton *et al.* (1989), reproduced with permission of the American Meteorological Society.]

be seen that the two measures usually rate a given forecast map similarly, although some differences are apparent. For example, in this data set the AC values in Fig. 7.14 show a more consistent separation between the performance of the dynamical and persistence forecasts than do the RMSE values in Fig. 7.13.

As was the case for the MSE, aggregate performance of a collection of field forecasts can be summarized by averaging anomaly correlations across many forecasts. However, skill scores of the form of Eq. (7.4) are seldom calculated for the AC, perhaps because AC_{Clim} is undefined [zero in the denominator of Eq. (7.39)]. Rather, skill with respect to the anomaly correlation is generally evaluated relative to the reference values AC = 0.6 or AC = 0.5. Individuals working operationally with the anomaly correlation have found, subjectively, that AC = 0.6 seems to represent a reasonable lower limit for delimiting field forecasts that are synoptically useful (Hollingsworth *et al.*, 1980). Murphy and Epstein (1989) have shown that if the field forecasts are unbiased, if the average forecast and observed anomalies are zero, and if the forecast field exhibits a realistic level of variability [i.e., the two summations in the denominator of Eq. (7.39) are of comparable magnitude], then AC = 0.5 corresponds to the skill score for the MSE in Eq. (7.38) being zero. Under these same restrictions, AC = 0.6 corresponds to the MSE skill score being 0.20.

Figure 7.15 illustrates the use of the subjective AC = 0.6 cutoff for useful forecast skill. Panel a shows average AC values for 500-mb-height forecasts made during the winters (December–February) of 1981/82 through 1989/90. For pro-

a
Anomaly Correlation (%)
Forecast Day
88/89
87/88
82/83
86/87
89/90
85/86
81/82
84/85
83/84

b
Time of Crossover (days)
Year
NMC (60% AC)
PERSIST. (40% AC)
PERSIST. (60% AC)

jection zero days into the future (i.e., initial time), AC = 1 since $y_m = o_m$ at all gridpoints. The average AC declines progressively for longer forecast lead times, falling below AC = 0.6 for projections between 5 and 7 days. The curves for the later years tend to lie above the curves for the earlier years, reflecting, at least in part, improvements made to the forecast model during this decade. One measure of this overall skill improvement is the increase in the average lead time at which the AC curve crosses the 0.6 line. These times are plotted in Fig. 7.15b, and range from 5 days in the early and mid-1980s, to 7 days in the late 1980s. Also plotted in this panel are the average projections at which anomaly correlations for persistence forecasts fall below 0.4 and 0.6. The increase for the AC = 0.4 threshold in later winters indicates that some of the apparent skill increase for the dynamical forecasts may be attributable to more persistent synoptic patterns. The crossover time at the AC = 0.6 threshold for persistence forecasts is consistently about 2 days. Thus, imagining the average correspondence between observed 500-mb maps separated by 48-h intervals allows a qualitative appreciation of the level of skill represented by the AC = 0.6 threshold.

Exercises

7.1. For the forecast verification data in Table 7.3:
   (a) Reconstruct the joint distribution, $p(y_i, o_j)$, $i = 1, \ldots, 12, j = 1, 2$.
   (b) Compute the unconditional (sample climatological) probability $p(o_1)$.

7.2. Construct the 2×2 contingency table that would result if the probability forecasts in Table 7.3 had been converted to categorical rain/no rain forecasts, with a threshold probability of 0.25.

7.3. Using the 2×2 contingency table from Exercise 7.2, compute
   (a) The hit rate
   (b) The threat score
   (c) The Heidke score
   (d) The Kuipers score

7.4. For the event o_3 (3–4 in. of snow) in Table 7.5, below, find
   (a) The threat score
   (b) The probability of detection
   (c) The false alarm rate
   (d) The bias ratio

Fig. 7.15 (a) Average anomaly correlations as a function of forecast projection for 1981/82–1989/90 winter 500-mb heights between 20° and 80° N. Accuracy decreases as forecast projection increases, but there are substantial differences between winters. (b) Average projection at which National Meteorological Center (NMC) forecast anomaly correlations cross the 0.6 level, and persistence forecasts cross the 0.4 and 0.6 levels, for Januarys and Februarys of these nine winters. [From Kalnay *et al.* (1990), reproduced with permission of the American Meteorological Society.]

Table 7.5

A 4 × 4 Contingency Table for Snow Amount Forecasts in Eastern Region of the United States during Winters 1983/84–1988/89[a]

	o_1	o_2	o_3	o_4
f_1	35,915	477	80	28
f_2	280	162	51	17
f_3	50	48	34	10
f_4	28	23	185	34

[a]The event o_1 is 0–1 in., o_2 is 2–3 in., o_3 is 3–4 in., and o_4 is $\geq$ 6 in. [From Goldsmith (1990).]

7.5. Using the 4 × 4 contingency table in Table 7.5, compute
   (a) The joint distribution of the forecasts and the observations
   (b) The hit rate
   (c) The Heidke score
   (d) The Kuipers score

7.6. Consider persistence forecasts for the January 1987 Ithaca maximum temperatures in Table A.1 (e.g., the forecast for January 2 is the observed temperature on January 1). Compute
   (a) The MAE
   (b) The RMSE
   (c) The ME (bias)
   (d) The skill, in terms of RMSE, with respect to the sample climatology.

7.7. Using the collection of hypothetical PoP forecasts summarized in Table 7.6:
   (a) Calculate the Brier score
   (b) Calculate the Brier score for (the sample) climatological forecast.
   (c) Calculate the skill of the forecasts with respect to the sample climatology.
   (d) Draw a reliability diagram for this set of forecasts.

Table 7.6

Hypothetical Verification Data for 1000 Probability-of-Precipitation Forecasts

Forecast probability, y_i	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Number of times forecast	293	237	162	98	64	36	39	26	21	14	10
Number of precipitation occurrences	9	21	34	31	25	18	23	18	17	12	9

Table 7.7

Hypothetical Verification for 500 Probability Forecasts of Precipitation Amounts

Forecast probabilities for			Number of forecast periods verifying as		
<0.01 in.	0.01–0.24 in.	≥0.25 in.	<0.01 in.	0.01–0.24 in.	≥0.25 in.
.8	.1	.1	263	24	37
.5	.4	.1	42	37	12
.4	.4	.2	14	16	10
.2	.6	.2	4	13	6
.2	.3	.5	4	6	12

7.8. Using the hypothetical probabilistic three-category precipitation amount forecasts in Table 7.7
   (a) Calculate the average RPS.
   (b) Calculate the skill of the forecasts with respect to the sample climatology.

7.9. For the hypothetical forecast and observed 500-mb fields in Fig. 7.16
   (a) Calculate the S1 score, comparing the 24 pairs of gradients in the north–south and east–west directions.
   (b) Calculate the MSE.
   (c) Calculate the skill score for the MSE with respect to the climatological field.
   (d) Calculate the AC.

Fig. 7.16 Hypothetical forecast (a), observed (b), and climatological average (c) fields of 500-mb heights (dam) over a small domain, and interpolations onto 16-point grids.

Chapter 8 Time Series

8.1 Background

This chapter presents methods for characterizing and analyzing the time variations of data series. Often one encounters data sets consisting of consecutive measurements of atmospheric variables. When the ordering of the data in time is important to their information content, summarization or analysis using time-series methods is appropriate.

As has been illustrated earlier, atmospheric observations separated by relatively short times tend to be similar, or correlated. Analyzing and characterizing the nature of these temporal correlations, or relationships through time, can be useful for both understanding atmospheric processes and for forecasting future atmospheric events. Accounting for these correlations is also necessary if valid statistical inferences about time-series data are to be made (Chapter 5).

8.1.1 Stationarity

Of course, one does not expect the future values of a data series to be identical to some past series of existing observations. However, in many instances it may be very reasonable to assume that their statistical properties will be the same. The idea that past and future values of a time series will be similar statistically is an informal expression of what is called *stationarity*. Usually, the term stationarity is understood to mean *weak stationarity*, or *covariance stationarity*. In this sense, stationarity implies that the mean and autocorrelation function [Eq. (3.24)] of the data series do not change through time. Different time slices of a stationary data series (e.g., the data observed to date and the data to be observed in the future) can be regarded as having the same underlying mean and variance. Furthermore, the correlation between variables in a stationary series is determined only by their separation in time [i.e., their lag, k, in Eq. (3.24)], and not their absolute positions in time. Qualitatively, different portions of a stationary time series "look alike," even though the individual data values may be very different. Covariance stationarity is a less restrictive assumption than *strict stationarity*, which implies that the full distribution of the variables in the series does not change through time.

More technical expositions of the concept of stationarity can be found in, for example, Fuller (1976) or Kendall and Ord (1990).

Most methods for analyzing time series assume stationarity of the data. However, many atmospheric processes are distinctly not stationary. Obvious examples of nonstationary atmospheric data series are those exhibiting annual or diurnal cycles. For example, temperatures typically exhibit very strong annual cycles in mid- and high-latitude climates, and one expects the average of the distribution of January temperature to be very different from that for July temperature. Similarly, time series of wind speeds often exhibit a diurnal cycle, which derives physically from the tendency for diurnal changes in static stability, imposing a diurnal cycle on downward momentum transport.

There are two approaches to dealing with nonstationary series. Both aim to process the data in a way that will subsequently allow stationarity to be reasonably assumed. The first approach is to mathematically transform the nonstationary data to approximate stationarity. For example, subtracting a periodic mean function from data subject to an annual cycle would produce a transformed data series with constant (zero) mean. In order to produce a series with both constant mean and variance, it might be necessary to further transform these anomalies to standardized anomalies [Eq. (3.16)]—specifically, to divide the values in the anomaly series by standard deviations that also vary through an annual cycle. Not only do temperatures tend to be colder in winter, but the variability of temperature tends to be higher. A possible approach to transforming monthly temperature series to (at least approximate) stationarity could be to compute the 12 monthly mean values and twelve monthly standard deviation values, and then to apply Eq. (3.16) using the different means and standard deviations for the appropriate calendar month. This was the first step used to construct the time series of SOI values in Fig. 3.10.

The alternative to data transformation is to stratify the data. That is, one can conduct separate analyses of subsets of the data record that are short enough to be regarded as nearly stationary. One might analyze daily observations for all available January data for a given location, assuming that each 31-day data record is a sample from the same physical process, but not necessarily assuming that process to be the same as for July, or even for February, data.

8.1.2 Time-Series Models

Characterization of the properties of a time series is often achieved by invoking mathematical models for the observed data variations. Having obtained a time-series model for an observed data set, that model might then be viewed as a *generating process*, or algorithm, that could have produced the data. A mathematical model for the time variations of a data set can allow compact representation of the characteristics of that data in terms of a few parameters. This approach is entirely analogous to the fitting of theoretical probability distributions, which constitute

another kind of probability model, in Chapter 4. The distinction is that the distributions in Chapter 4 are used without regard to the ordering of the data, whereas the motivation for using time-series models is to characterize the nature of the ordering. Time-series methods are thus appropriate when the ordering of the data values in time are important to a given application.

Regarding an observed time series as having been generated by a theoretical (model) process is convenient because it allows characteristics of future, yet unobserved, values of a time series to be inferred from the inevitably limited data at hand. That is, characteristics of an observed time series are summarized by the parameters of the time-series model. Invoking the assumption of stationarity, future values of the time series should then also exhibit the statistical properties implied by the model. The properties of the model-generating process can be used to infer characteristics of yet unobserved values of the series.

8.1.3 Time-Domain versus Frequency-Domain Approaches

There are two fundamental approaches to time-series analysis: time-domain analysis and frequency-domain analysis. These two approaches proceed very differently and may seem quite distinct, but they are not independent. Rather, they are complementary methods that are linked mathematically.

Time-domain methods seek to characterize data series in the same terms in which they are observed and reported. A primary tool for characterization of relationships between data values in the time-domain approach is the autocorrelation function. Mathematically, time-domain analyses operate in the same space as the data values. Separate sections in this chapter describes different time-domain methods for use with discrete and continuous data. Here the terms *discrete* and *continuous* are used in the same sense as in Chapter 4: Discrete random variables are allowed to take on only a finite (or possibly countably infinite) number of values, whereas continuous random variables may take on any of the infinitely many real values within their range.

Frequency-domain analysis represents data series in terms of contributions occurring at different time scales, or characteristic frequencies. Each time scale is represented by a pair of sine and cosine functions. The overall time series is regarded as having arisen from the combined effects of a collection of sine and cosine waves oscillating at different rates. The sum of these waves reproduces the original data, but it is often the relative strengths of the individual component waves that are of primary interest. Frequency-domain analyses take place in the mathematical space defined by this collection of sine and cosine waves. That is, frequency-domain analysis involves transformation of the n original data values into coefficients that multiply an equal number of periodic (the sine and cosine) functions. At first exposure this process can seem very strange, and is sometimes difficult to grasp. However, frequency-domain methods are very commonly applied to atmospheric time series, and important insights can be gained from frequency-domain analyses.

8.2 Time Domain. I. Discrete Data

8.2.1 Markov Chains

Recall that a discrete random variable is one that can take on values only from among a defined, finite set. The most common class of model, or stochastic process, used to represent time series of discrete variables is known as the *Markov chain*. A Markov chain can be imagined as being based on collection of "states" of a model system. Each state corresponds to one of the elements of the MECE partition of the sample space comprising the random variable in question.

For each time period, the length of which is equal to the time separation between observations in the time series, the Markov chain can either remain in the same state or change to one of the other states. Remaining in the same state corresponds to two successive observations of the same value of the random variable in the time series, whereas a change of state implies two successive values of the time series that are different.

The behavior of a Markov chain is governed by a set of probabilities for these transitions, called the *transition probabilities*. The transition probabilities specify likelihoods for the system being in each of its possible states during the next time period. The most common form is called a *first-order Markov chain*, for which the transition probabilities controlling the next state of the system depend only on the current state of the system. That is, knowledge of the current state of the system and the full sequence of states leading up to the current state provides no more information about the probability distribution for the states at the next observation time than does knowledge of the current state alone. This characteristic of first-order Markov chains is known as the *Markovian property*, which can be expressed more formally as

$$\Pr\{X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \ldots, X_1\} = \Pr\{X_{t+1} \mid X_t\}. \tag{8.1}$$

The probabilities of future states depend on the current state, but they do not depend on the particular way that the model system arrived at the current state. In terms of a time series of observed data the Markovian property means, for example, that forecasts of tomorrow's data value can be made on the basis of today's observation, but also that knowing yesterday's data value provides no additional information.

The transition probabilities of a Markov chain are conditional probabilities. For a first-order Markov chain, there is a conditional probability distribution pertaining to each possible current state of the system, and each of these distributions specifies probabilities for the states of the system in the next time period. To say that these probability distributions are conditional allows for the possibility that the transition probabilities can be different, depending on the current state. The fact that these distributions can be different is the essence of the ability of a Markov chain to represent the serial correlation, or persistence, often exhibited by atmospheric variables. If the probability of the future state is the same, regardless

of the current state, then the time series consists of independent values. In that case the probability of occurrence of any given state in the upcoming time period is not affected by the occurrence or nonoccurrence of a particular state in the current time period. If the time series being modeled exhibits persistence, the probability of the system staying in a given state will tend to be higher than the probabilities of arriving at that state from other states.

If the transition probabilities of a Markov chain do not change through time and none of them are zero, then the resulting time series will be stationary. Modeling nonstationary data series exhibiting, for example, an annual cycle can require allowing the transition probabilities to vary through an annual cycle as well. One way to achieve this is to specify that the probabilities vary according to some smooth periodic curve, such as a cosine function. Alternatively, separate transition probabilities can be used for nearly stationary portions of the cycle, for example four 3-month seasons or 12 calendar months.

Certain classes of Markov chains are described more concretely, but relatively informally, in the following sections. More formal and comprehensive treatments can be found in, for example, Feller (1970), Karlin and Taylor (1975), or Katz (1985).

8.2.2 Two-State, First-Order Markov Chains

The simplest kind of discrete random variable pertains to the situation of dichotomous (yes/no) events. The behavior of a stationary sequence of independent (exhibiting no serial correlation) values of a dichotomous discrete random variable is described by the binomial distribution [Eq. (4.1)]. That is, for serially independent events, the ordering in time is of no importance from the perspective of specifying probabilities for future events, so that a time-series model for their behavior does not provide more information than does the simple binomial distribution.

A two-state Markov chain is a statistical model for the persistence of binary events. The occurrence or nonoccurrence of rain on a given day is a simple meteorological example of a binary random event, and a sequence of daily observations of "rain" and "no rain" for a particular location would constitute a time series of that variable. Consider a series where the random variable takes on the values $X_t = 1$ if precipitation occurs on day t and $X_t = 0$ if it does not. For the January 1987 Ithaca precipitation data in Table A.1, this time series would consist of the values shown in Table 8.1. That is, $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, $x_4 = 0$, $\ldots$, and $x_{31} = 1$. It is evident from looking at this series of numbers that the 1's and 0's tend to cluster in time. As was illustrated in Example 2.2, this clustering is an expression of the serial dependence present in the time series. That is, the probability of a "1" following a "1" is apparently higher than the probability of a "1" following a "0," and the probability of a "0" following a "0" is apparently higher than the probability of a "0" following a "1."

A common and often quite good stochastic model for data of this kind is a first-order, two-state Markov chain. A two-state Markov chain is natural for

Table 8.1

Time Series of a Dichotomous Random Variable Derived from the January 1987 Ithaca Precipitation Data in Table A.1[a]

Date, t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
X_t	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	
Date, t	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31		
X_t	0	0	0	1	0	0	1	0	0	0	0	0	1	1	1		

[a] Days on which nonzero precipitation was reported yield $X_t = 1$; days with zero precipitation yield $X_t = 0$.

dichotomous data since each of the two states will pertain to one of the possible data values. A first-order Markov chain has the property that the transition probabilities governing each observation in the time series depends only on the value of the previous member of the time series.

Figure 8.1 illustrates schematically the nature of a two-state first-order Markov chain. In order to help fix ideas, the two states are labeled in a manner consistent with the data in Table 8.1. For each value of the time series, the stochastic process is either in state 0 (no precipitation occurs and $X_t = 0$), or in state 1 (precipitation occurs and $X_t = 1$). At each time step the process can either stay in the same state or switch to the other state. Therefore four distinct transitions are possible, corresponding to a dry day following a dry day (p_{00}), a wet day following a dry day (p_{01}), a dry day following a wet day (p_{10}), and a wet day following a wet day (p_{11}). Each of these four transitions is represented in Fig. 8.1 by arrows, labeled with the appropriate transition probabilities. Here the notation is such that the first subscript on the probability is the state at time t, and the second subscript is the state at time $t + 1$.

The transition probabilities are conditional probabilities for the state at time

Fig. 8.1 Schematic representation of a two-state, first-order Markov chain, illustrated in terms of daily precipitation occurrence or nonoccurrence. The two states are labeled "0" for no precipitation, and "1" for precipitation occurrence. For a first-order Markov chain, there are four transition probabilities controlling the state of the system in the next time period. Since these four probabilities are pairs of conditional probabilities, $p_{00} + p_{01} = 1$ and $p_{10} + p_{11} = 1$. For quantities such as day-to-day precipitation occurrence that exhibit positive serial correlation, $p_{10} < p_{00}$, and $p_{01} < p_{11}$.

$t + 1$ (e.g., whether precipitation will occur tomorrow) given the state at time t (e.g., whether precipitation occurred today). That is,

$$p_{00} = \Pr\{X_{t+1} = 0 \mid X_t = 0\} \tag{8.2a}$$

$$p_{01} = \Pr\{X_{t+1} = 1 \mid X_t = 0\} \tag{8.2b}$$

$$p_{10} = \Pr\{X_{t+1} = 0 \mid X_t = 1\} \tag{8.2c}$$

$$p_{11} = \Pr\{X_{t+1} = 1 \mid X_t = 1\}. \tag{8.2d}$$

Together, Eqs. (8.2a) and (8.2b) constitute the conditional probability distribution for the value of the time series at time $t + 1$, given that $X_t = 0$ at time t. Similarly, Eqs. (8.2c) and (8.2d) express the conditional probability distribution for the next value of the time series given that the current value is $X_t = 1$.

Notice that the four probabilities in Eqs. (8.2) provide some redundant information. Given that the Markov chain is in one state or the other at time t, the sample space for X_{t+1} consists of only two MECE events. Therefore, $p_{00} + p_{01} = 1$ and $p_{10} + p_{11} = 1$, so that it is really necessary to focus on only one of each of the pairs of transition probabilities, say, p_{01} and p_{11}. In particular, it is sufficient to estimate only two parameters for a two-state first-order Markov chain, since the two pairs of conditional probabilities must sum to 1. The parameter estimation procedure consists simply of computing the conditional relative frequencies, which yield the maximum-likelihood estimators (MLEs)

$$\hat{p}_{01} = \frac{\text{\# of 1's following 0's}}{\text{total \# of 0's}} = \frac{n_{01}}{n_{0\cdot}} \tag{8.3a}$$

and

$$\hat{p}_{11} = \frac{\text{\# of 1's following 1's}}{\text{total \# of 1's}} = \frac{n_{11}}{n_{1\cdot}}. \tag{8.3b}$$

Here n_{01} is the number of transitions from state 0 to state 1, n_{11} is the number of pairs of time steps in which there are two consecutive 1's in the series, $n_{0\cdot}$ is the number of 0's in the series followed by another data point, and $n_{1\cdot}$ is the number of 1's in the series followed by another data point. That is, the subscript "·" indicates the total over all values of the index replaced by this symbol, so that $n_{1\cdot} = n_{10} + n_{11}$ and $n_{0\cdot} = n_{00} + n_{01}$. Equations (8.3) state that the parameter p_{01} is estimated by looking at the conditional relative frequency of the event $x_{t+1} = 1$ considering only those points in the time series following data values for which $x_t = 0$. Similarly, p_{11} is estimated as that fraction of points for which $x_t = 1$ is followed by points with $x_{t+1} = 1$. The somewhat labored definitions of $n_{0\cdot}$ and $n_{1\cdot}$ are necessary to account for the "edge" effect of a finite sample size. The final point in the time series is not counted in the denominator of Eq. (8.3a) or (8.3b), whichever is appropriate, because there is no available data value following it to be incorporated into the counts in one of the numerators. These definitions also

cover cases of missing values, and stratified samples such as 30 years of January data, for example.

Equation (8.3) suggests that parameter estimation for a two-state first-order Markov chain is equivalent to fitting two Bernoulli distributions (i.e., binomial distributions with $N = 1$). One of these binomial distributions pertains to points in the time series preceded by 0's, and the other describes the behavior of points in the time series preceded by 1's. Knowing that the process is currently in state 0 (e.g., no precipitation today), the probability distribution for the event $X_{t+1} = 1$ (precipitation tomorrow) is simply binomial [Eq. (4.1)] with $p = p_{01}$. The second binomial parameter is $N = 1$, because there is only one data point in the series for each time step. Similarly, if $X_t = 1$, then the distribution for the event $X_{t+1} = 1$ is binomial with $N = 1$ and $p = p_{11}$. The conditional dichotomous events of a stationary Markov chain satisfy the requirements listed in Chapter 4 for the binomial distribution. For a stationary process the probabilities do not change through time, and conditioning on the current value of the time series satisfies the independence assumption for the binomial distribution because of the Markovian property. It is the fitting of two binomial distributions that allows the time dependence in the data series to be represented.

Certain properties are implied for a time series described by a Markov chain. These properties are controlled by the values of the transition probabilities, and can be computed from them. First, the long-run relative frequencies of the events corresponding to the two states of the Markov chain are called the *stationary probabilities*. For a Markov chain describing the daily occurrence or nonoccurrence of precipitation, the stationary probability for precipitation, π_1, corresponds to the (unconditional) climatological probability of precipitation. In terms of the transition probabilities p_{01} and p_{11},

$$\pi_1 = \frac{p_{01}}{1 + p_{01} - p_{11}}, \tag{8.4}$$

with the stationary probability for state 0 being simply $\pi_0 = 1 - \pi_1$. In the usual situation of positive serial correlation or persistence, one finds $p_{01} < \pi_1 < p_{11}$. Applied to daily precipitation occurrence, this means that the conditional probability of a wet day following a dry day is less than the overall climatological relative frequency, whereas the conditional probability of a wet day following a wet day is greater than the climatological probability.

The transition probabilities also imply a degree of serial correlation, or persistence, for the time series. In terms of the transition probabilities, the lag-1 autocorrelation [Eq. (3.23)] of the time series is simply

$$r_1 = p_{11} - p_{01}. \tag{8.5}$$

In the context of Markov chains, r_1 is sometimes known as the *persistence parameter*. As the correlation r_1 increases, the difference between p_{11} and p_{01} widens. This implies that state 1 is more and more likely to follow state 1, and less

and less likely to follow state 0. That is, there is an increasing tendency for 0's and 1's to cluster in time, or occur in runs. A time series exhibiting no autocorrelation would be characterized by $r_1 = p_{11} - p_{01} = 0$, or $p_{11} = p_{01} = \pi_1$. In this case the two conditional probability distributions specified by Eq. (8.2) are the same, and the time series is simply a string of independent binomial realizations, each with $N = 1$. The binomial distribution with $N = 1$ can be viewed as a two-state, *zero*-order Markov chain.

The full autocorrelation function, Eq. (3.24), follows easily from the lag-1 correlation r_1. Because of the Markovian property, the autocorrelation between members of the time series separated by k time steps is simply the lag-1 autocorrelation multiplied by itself k times,

$$r_k = (r_1)^k. \tag{8.6}$$

A common misconception is that the Markovian property implies independence of values in a first-order Markov chain that are separated by more than one time period. Equation (8.6) shows that the correlation, and hence the statistical dependence, among elements of the time series tails off at increasing lags, but it is never exactly zero unless $r_1 = 0$. Rather, the Markovian property implies *conditional* independence of data values separated by more than one time period, as expressed by Eq. (8.1). Given a particular value for X_t, the different possible values for X_{t-1}, X_{t-2}, X_{t-3}, and so on do not affect the probabilities for X_{t+1}. However, for example, $\Pr\{X_{t+1} = 1 \mid X_{t-1} = 1\} \neq \Pr\{X_{t+1} = 1 \mid X_{t-1} = 0\}$, indicating statistical dependence among members of a Markov chain separated by more than one time period.

8.2.3 Test for Independence versus First-Order Serial Dependence

Even if a series of binary data is generated by a mechanism producing serially independent values, the sample lag-1 autocorrelation [Eq. (8.5)] computed from a finite sample is unlikely to be exactly zero. A formal test, similar to the χ^2 goodness-of-fit test [Eq. (5.18)], can be conducted to investigate the statistical significance of the sample autocorrelation for a binary data series. The null hypothesis for this test is that the data series is serially independent (i.e., the data are independent binomial variables with $N = 1$); the alternative is that the series was generated by a first-order Markov chain.

The test is based on a contingency table of the observed transition counts n_{00}, n_{01}, n_{10}, and n_{11}, in relation to the numbers of transitions expected under the null hypothesis. The corresponding expected counts, e_{00}, e_{01}, e_{10}, and e_{11}, are computed from the observed transition counts under the constraint that the marginal totals of the expected counts are the same as for the observed transitions. The situation is illustrated in Fig. 8.2, which shows generic contingency tables for the observed transition counts (a) and those expected under the null hypothesis of independence (b). For example, the transition count n_{00} specifies the number of

(a)

	$X_{t+1} = 0$	$X_{t+1} = 1$	
$X_t = 0$	$\mathbf{n_{00}}$	$\mathbf{n_{01}}$	$n_{0\bullet}$
$X_t = 1$	$\mathbf{n_{10}}$	$\mathbf{n_{11}}$	$n_{1\bullet}$
	$n_{\bullet 0}$	$n_{\bullet 1}$	n

(b)

	$X_{t+1} = 0$	$X_{t+1} = 1$	
$X_t = 0$	$\mathbf{e_{00}} = (n_{0\bullet})(n_{\bullet 0})/n$	$\mathbf{e_{01}} = (n_{0\bullet})(n_{\bullet 1})/n$	$n_{0\bullet}$
$X_t = 1$	$\mathbf{e_{10}} = (n_{1\bullet})(n_{\bullet 0})/n$	$\mathbf{e_{11}} = (n_{1\bullet})(n_{\bullet 1})/n$	$n_{1\bullet}$
	$n_{\bullet 0}$	$n_{\bullet 1}$	n

Fig. 8.2 Contingency tables of observed transition counts n_{ij} (a) for a binary time series and (*b*) transition counts e_{ij} expected if the time series actually consists of serially independent values with the same marginal totals. The transition counts are shown in boldface, and the marginal totals are in plain type.

consecutive pairs of 0's in the time series. This is related to the joint probability $\Pr\{X_t = 0 \cap X_{t+1} = 0\}$. Under the null hypothesis of independence this joint probability is simply the product of the two event probabilities, or in relative frequency terms, $\Pr\{X_t = 0\} \Pr\{X_{t+1} = 0\} = (n_{0\cdot}/n)(n_{\cdot 0}/n)$. Thus, the corresponding number of expected transition counts is simply this product multiplied by the sample size, or $e_{00} = (n_{0\cdot})(n_{\cdot 0})/n$. More generally

$$e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}. \tag{8.7}$$

The test statistic is computed from the observed and expected transition counts using

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{8.8}$$

where, for the 2×2 contingency table appropriate for dichotomous data, the summations are for $i = 0$ to 1 and $j = 0$ to 1. That is, there is a separate term in Eq. (8.8) for each of the four pairs of contingency table cells in Fig. 8.2. Note that Eq. (8.8) is analogous to Eq. (5.18), with the n_{ij} being the observed counts, and the e_{ij} being the expected counts. Under the null hypothesis, the test statistic follows the χ^2 distribution with the parameter $\nu = 1$ degree of freedom. This value of the degrees-of-freedom parameter is appropriate because, given that the marginal totals are fixed, arbitrarily specifying one of the transition counts completely determines the other three.

The fact that the numerator in Eq. (8.8) is squared implies that values of the test statistic on the left tail of the null distribution are favorable to H_0, because small values of the test statistic are produced by pairs of observed and expected transition counts of similar magnitudes. Therefore, the test is one-tailed. The p value

associated with a particular test can be assessed using tables of the χ^2 distribution, or by using Table B.2 (in Appendix B) since the χ^2 is a special case of the gamma distribution. Note that the test statistic must be transformed to the standard gamma distribution (with $\beta = 1$), as described in Chapter 5 and illustrated in the following example.

Example 8.1. Fitting a Two-State, First-Order Markov Chain

Consider summarizing the time series in Table 8.1, derived from the January 1987 Ithaca precipitation series in Table A.1, using a first-order Markov chain. The parameter estimates in Eq. (8.3) are easily obtained from the transition counts. For example, the number of 1's following 0's in the time series of Table 8.1 is $n_{01} = 5$. Similarly, $n_{00} = 11$, $n_{10} = 4$, and $n_{11} = 10$. The resulting sample estimates for the transition probabilities [Eqs. (8.3)] are $p_{01} = 5/16 = 0.312$, and $p_{11} = 10/14 = 0.714$. Note that these are identical to the conditional probabilities computed in Example 2.2.

Whether the extra effort of fitting the first-order Markov chain to the data in Table 8.1 is justified can be investigated using the χ^2 test in Eq. (8.8). Here the null hypothesis is that these data resulted from an independent (i.e., binomial) process, and the expected transition counts e_{ij} that must be computed are those consistent with this null hypothesis. These are obtained from the marginal totals $n_{0\cdot} = 11 + 5 = 16$, $n_{1\cdot} = 4 + 10 = 14$, $n_{\cdot 0} = 11 + 4 = 15$, and $n_{\cdot 1} = 5 + 10 = 15$. The expected transition counts follow easily as $e_{00} = (16)(15)/30 = 8$, $e_{01} = (16)(15)/30 = 8$, $e_{10} = (14)(15)/30 = 7$, and $e_{11} = (14)(15)/30 = 7$. Note that usually the expected transition counts will be different from each other, and need not be integer values.

Computing the test statistic in Eq. (8.8), one finds $\chi^2 = (11 - 8)^2/8 + (5 - 8)^2/8 + (4 - 7)^2/7 + (10 - 7)^2/7 = 4.82$. The degree of unusualness of this result with reference to the null hypothesis can be assessed with the aid of the gamma distribution probabilities in Table B.2. Recall that the χ^2 distribution is a special case of the gamma distribution, with $\alpha = \nu/2 = 0.5$, and $\beta = 2$. Transforming to the standard (i.e., $\beta = 1$) gamma distribution using $\xi = \chi^2/\beta = 4.82/2 = 2.41$, one finds that the result lies between the 95th and 99th percentiles of the standard gamma distribution with $\alpha = 0.5$. Thus, even for this rather small sample size, the null hypothesis of serial independence would be rejected at the 5% level, although not at the 1% level.

The degree of persistence exhibited by this data sample can be summarized using the persistence parameter, which is also the lag-1 autocorrelation, $r_1 = p_{11} - p_{01} = 0.714 - 0.312 = 0.402$. This is a fairly large value, indicative of substantial serial correlation in the time series. It also implies the full autocorrelation function, through Eq. (8.6). Figure 8.3 shows that the implied theoretical correlation function for this Markov process, shown as the dashed line, agrees very closely with the sample autocorrelation function shown by the solid line, for the first few lags. This agreement provides qualitative support for the first-order Markov chain as an appropriate model for the data series.

Fig. 8.3 Sample autocorrelation function for the January 1987 Ithaca binary precipitation occurrence series, Table 8.1 (solid, with circles), and theoretical autocorrelation function (dashed) specified by the fitted first-order Markov chain model [Eq. (8.6)]. The correlations are 1.00 for $k = 0$, since the unlagged data are perfectly correlated with themselves.

Finally, the stationary (i.e., climatological) probability for precipitation implied for this data by the Markov chain model is, using Eq. (8.4), $\pi_1 = 0.312/(1 + 0.312 - 0.714) = 0.522$. This value agrees closely with the relative frequency $16/30 = 0.533$, obtained by counting the number of 1's in the last thirty values of the series in Table 8.1. ¤

8.2.4 Some Applications of Two-State Markov Chains

One interesting application of the Markov chain model is in the computer generation of synthetic rainfall series. Time series of random binary numbers, statistically resembling real rainfall occurrence data, can be generated using the Markov chain as an algorithm. To generate sequences of numbers statistically resembling those in Table 8.1, for example, the parameters $p_{01} = 0.312$ and $p_{11} = 0.714$, estimated in Example 8.1, would be used together with a uniform [0,1] random-number generator. The uniform random-number generator is a computer algorithm producing numbers in the interval [0,1] with equal probability. The synthetic time series would begin using the stationary probability $\pi_1 = 0.522$. If the first uniform number generated were less than π_1, then $x_1 = 1$, meaning that the first simulated day would be wet. For subsequent values in the series, each new uni-

form random number is compared to the appropriate transition probability, depending on whether the most recently generated number, corresponding to day t, was wet or dry. That is, the transition probability p_{01} would be used to generate X_{t+1} if $X_t = 0$, and p_{11} would be used if $X_t = 1$. A wet day ($X_{t+1} = 1$) is simulated if the next uniform random number is less than the transition probability, and a dry day ($X_{t+1} = 0$) is generated if it is not. Since typically $p_{11} > p_{01}$ for precipitation data, simulated wet days are more likely to follow wet days than dry days, as occurs in the real data series.

The Markov chain approach for simulating precipitation occurrence can be extended to include simulation of daily precipitation amounts. This is accomplished by adopting a statistical model for the nonzero rainfall amounts, yielding a sequence of random variables defined on the Markov chain, called a *chain-dependent process* (Katz, 1977). Commonly a gamma distribution (Chapter 4) is fit to the precipitation amounts on wet days in the data record (e.g., Katz, 1977; Richardson, 1981; Stern and Coe, 1984), although other distributions can also be used (Woolhiser and Roldan, 1982). Computer algorithms are available to generate random variables drawn from gamma distributions (e.g., Bratley *et al.*, 1983; Johnson, 1987), and these can be used to simulate precipitation amounts on days when the Markov chain calls for a wet day. The tacit assumption that precipitation amounts on consecutive wet days are independent has turned out to be a reasonable approximation in those instances where it has been investigated (Katz, 1977; Stern and Coe, 1984). Generally both the Markov chain transition probabilities and the parameters of the distributions describing precipitation amounts change through the year. These seasonal cycles can be handled by fitting separate sets of parameters for each of the 12 calendar months (e.g., Wilks 1989), or by representing them using smoothly varying sine and cosine functions (Stern and Coe, 1984).

Inferences concerning properties of longer-term precipitation quantities (e.g., the monthly frequency distributions of numbers of wet days in a month, or of total monthly precipitation) can be drawn from the parameters of the chain-dependent process that governs the daily precipitation series. Since observed monthly precipitation statistics are computed from individual daily values, it should not be surprising that the statistical characteristics of monthly precipitation quantities will depend directly on the statistical characteristics of daily precipitation occurrence and amount. Katz (1977, 1985) gives equations specifying some of these relationships, which can be used in a variety of ways (e.g., Katz and Parlange, 1993; Wilks, 1992).

Finally, another interesting perspective on the Markov chain model for daily precipitation occurrence is in relation to forecasting precipitation probabilities. Recall that forecast skill is assessed relative to a set of benchmark, or reference forecasts (Chapter 7). Usually one of two reference forecasts is used: (1) the climatological probability of the forecast event, in this case π_1; or (2) persistence forecasts specifying unit probability if precipitation occurred in the previous period, or zero probability if the event did not occur. Neither of these reference

forecasting systems is particularly sophisticated, and both are relatively easy to improve on, at least for short-range forecasts. A more challenging, yet still fairly simple alternative is to use the transition probabilities of a two-state Markov chain as the reference forecasts. If precipitation did not occur in the preceding period, the reference forecast would be p_{01}, whereas the conditional forecast probability for precipitation following a day with precipitation would be p_{11}. Note that for meteorological quantities exhibiting persistence, $0 < p_{01} < \pi_1 < p_{11} < 1$, so that reference forecasts consisting of Markov chain transition probabilities constitute a compromise between persistence (either 0 or 1) and climatology (π_1). Furthermore, the balance of this compromise depends on the strength of the persistence exhibited by the climatological data on which the estimated transition probabilities are based. A weakly persistent quantity would be characterized by transition probabilities differing little from π_1, whereas a strong serial correlation will produce transition probabilities much closer to 0 and 1.

8.2.5 Multistate Markov Chains

Markov chains are also useful for representing the time correlation of discrete variables that can take on more than two values. For example, a three-state, first-order Markov chain is illustrated schematically in Fig. 8.4. Here the three states are arbitrarily labeled 1, 2, and 3. At each time t, the random variable in the series

Fig. 8.4 Schematic illustration of a three-state, first-order Markov chain. There are nine possible transitions among the three states, including the possibility that two consecutive points in the time series will be in the same state. First-order time dependence implies that the transition probabilities depend only on the current state of the system, or present value of the time series.

can take on one of the three values $X_t = 1$, $X_t = 2$, or $X_t = 3$, and each of these values corresponds to a state. First-order time dependence implies that the transition probabilities for X_{t+1} depend only on the state X_t, so that there are $3^2 =$ nine transition probabilities, p_{ij}. In general, for a first-order, k-state Markov chain, there are k^2 transition probabilities.

As is the case for the two-state Markov chain discussed above, the transition probabilities for multiple-state Markov chains are conditional probabilities. For example, the transition probability p_{12} in Fig. 8.4 is the conditional probability that state 2 will occur at time $t + 1$, given that state 1 occurred at time t. Therefore, in a k-state Markov chain it must be the case that the probabilities for the k transitions emanating from each state must sum to one, or $\Sigma_j p_{ij} = 1$ for each value of i.

Estimation of the transition probabilities for multiple-state Markov chains is a straightforward generalization of the formulas in Eqs. (8.3) for two-state chains. Each of these estimates is simply obtained from the conditional relative frequencies of the transition counts:

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{i\cdot}}; \qquad i, j = 1, \ldots, s. \tag{8.9}$$

As before, the dot indicates summation over all values of the replaced subscript so that, for example, $n_{1\cdot} = \Sigma_j n_{1j}$. For the $s = 3$-state Markov chain represented in Fig. 8.4, for example, $\hat{p}_{12} = n_{12}/(n_{11} + n_{12} + n_{13})$. In general, a contingency table of transition counts, corresponding to Fig. 8.2a for the $s = 2$-state case, will contain s^2 entries.

Testing whether the observed degree of serial correlation is significantly different from zero in a multiple-state situation can be done using the χ^2 test in Eq. (8.8). Here the summations are over all s possible states, and will include s^2 terms. As before the expected numbers of transition counts e_{ij} are computed using Eq. (8.7). Under the null hypothesis of no serial correlation, the distribution of the test statistic in Eq. (8.8) is χ^2 with $\nu = (s - 1)^2$ degrees of freedom.

Three-state Markov chains have been used to characterize transitions between below-normal, near-normal, and above-normal months for temperature and precipitation, as defined by the U.S. Climate Analysis Center (Examples 4.6 and 4.8), by Preisendorfer and Mobley (1984) and Wilks (1989). Mo and Ghil (1987) used a five-state Markov chain to characterize transitions between persistent hemispheric 500-mb flow types.

8.2.6 Higher-Order Markov Chains

First-order Markov chains often provide good representations of daily precipitation occurrence; however, it is not obvious just from inspection of the series in Table 8.1, for example, that this simple model will be adequate to capture the exhibited correlation structure. More generally, an m-order Markov chain is one

where the transition probabilities depend on the states in the previous m time periods. Formally, the extension of the Markovian property expressed in Eq. (8.1) to the m-order Markov chain is

$$\Pr\{X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \ldots, X_1\} = \Pr\{X_{t+1} \mid X_t, X_{t-1}, \ldots, X_{t-m}\}. \quad (8.10)$$

Consider, for example, a second-order Markov chain. Second-order time dependence means that the transition probabilities depend on the states (values of the time series) at lags of both one and two time periods. Notationally, then, the transition probabilities for a second-order Markov chain require three subscripts: the first denoting the state at time $t - 1$, the second denoting the state in time t, and the third specifying the state at (the future) time $t + 1$. The notation for the transition probabilities of a second-order Markov chain can be defined as

$$p_{hij} = \{X_{t+1} = j \mid X_t = i, X_{t-1} = h\}. \quad (8.11)$$

In general, the notation for a m-order Markov chain requires $m + 1$ subscripts on the transition counts and transition probabilities. If Eq. (8.11) is being applied to a binary time series such as that in Table 8.1, the model would be a two-state, second-order Markov chain, and the indices h, i, and j could take on either of the $s = 2$ values of the time series, say, 0 and 1. However, Eq. (8.11) is equally applicable to discrete time series with larger numbers ($s > 2$) of states.

As is the case for first-order Markov chains, transition probability estimates are obtained from relative frequencies of observed transition counts. However, since data values further back in time now need to be considered, the number of possible transitions increases exponentially with the order, m, of the Markov chain. In particular, for an s-state, m-order Markov chain, there are $s^{(m+1)}$ distinct transition counts and transition probabilities. The arrangement of the resulting transition counts, in the form of Fig. 8.2a, is shown in Table 8.2 for a $s = 2$ state, $m =$ 2-order Markov chain. The transition counts are determined from the observed

Table 8.2

Arrangement of the $2^{2+1} = 8$ Transition Counts for a Two-State, Second-Order Markov Chain in a Table of the Form of Fig. 8.2a[a]

X_{t-1}	X_t	$X_{t+1} = 0$	$X_{t+1} = 1$	Marginal totals
0	0	$\mathbf{n_{000}}$	$\mathbf{n_{001}}$	$n_{00\cdot} = n_{000} + n_{001}$
0	1	$\mathbf{n_{010}}$	$\mathbf{n_{011}}$	$n_{01\cdot} = n_{010} + n_{011}$
1	0	$\mathbf{n_{100}}$	$\mathbf{n_{101}}$	$n_{10\cdot} = n_{100} + n_{101}$
1	1	$\mathbf{n_{110}}$	$\mathbf{n_{111}}$	$n_{11\cdot} = n_{110} + n_{111}$

[a] Determining these counts from an observed time series requires examination of successive triplets of data values.

data series by examining consecutive groups of $m + 1$ data points. For example, the first three data points in Table 8.1 are $X_{t-1} = 0$, $X_t = 1$, $X_{t+1} = 1$, and this triplet would contribute one to the transition count n_{011}. Overall the data series in Table 8.1 exhibits three transitions of this kind, so the final transition count $n_{011} = 3$ for this data set. The second triplet in the data set in Table 8.1 would contribute one count to n_{110}. There is only one other triplet in this data for which $X_{t-1} = 1$, $X_t = 1$, $X_{t+1} = 0$, so the final count for $n_{110} = 2$.

The transition probabilities for a second-order Markov chain are obtained from the conditional relative frequencies of the transition counts

$$\hat{p}_{hij} = \frac{n_{hij}}{n_{hi\bullet}}. \tag{8.12}$$

That is, given that the value of the time series at time $t - 1$ was $X_{t-1} = h$ and the value of the time series at time t was $X_t = i$, the probability that the future value of the time series $X_{t+1} = j$ is p_{hij}, and the sample estimate of this probability is given in Eq. (8.12). Just as the two-state first-order Markov chain consists essentially of two conditional binomial distributions, a two-state second-order Markov chain amounts to four conditional binomial distributions, with parameters $N = 1$ and $p = p_{hi1}$, for each of the four distinct combinations of the indices h and i.

Note that the small data set in Table 8.1 is really too short to fit a second-order Markov chain. Since there are no triplets in this series for which $X_{t-1} = 1$, $X_t = 0$, $X_{t+1} = 1$ (i.e., a single dry day following and followed by a wet day) the transition count $n_{101} = 0$. This would lead to the sample estimate for the transition probability $\hat{p}_{101} = 0$, even though there is no physical reason why this particular sequence of wet and dry days should not occur.

8.2.7 Deciding among Alternative Orders of Markov Chains

How is one to know what order m is appropriate for a Markov chain to represent a particular data series? One approach is to use a hypothesis test. For example, the χ^2 test in Eq. (8.8) can be used to assess the significance of a first-order Markov chain model versus a zero-order, or binomial model (Example 8.1). While the mathematical structure of this test can be modified to investigate the suitability of, say, a first-order versus a second-order, or a second-order versus a third-order Markov chain, the overall significance of a collection of such tests would be difficult to evaluate. This difficulty arises in part because of the issue multiplicity. As discussed in the context of the topic of field significance in Chapter 5, the overall significance level of a collection of simultaneous, correlated tests is difficult to evaluate.

Two criteria are in common use for choosing among alternative orders of Markov chain models: the *Akaike information criterion* (AIC) (Akaike, 1974; Tong, 1975) and the *Bayesian information criterion* (BIC) (Schwarz, 1978; Katz, 1981). Both are based on the log-likelihood functions for the transition

probabilities of the fitted Markov chains. These log-likelihoods depend on the transition counts and the estimated transition probabilities. The log-likelihoods for Markov chains of order 0, 1, 2, and 3 are

$$L_0 = \sum_{j=0}^{s-1} n_j \ln(\hat{p}_j) \tag{8.13a}$$

$$L_1 = \sum_{i=0}^{s-1} \sum_{j=0}^{s-1} n_{ij} \ln(\hat{p}_{ij}) \tag{8.13b}$$

$$L_2 = \sum_{h=0}^{s-1} \sum_{i=0}^{s-1} \sum_{j=0}^{s-1} n_{hij} \ln(\hat{p}_{hij}) \tag{8.13c}$$

and

$$L_3 = \sum_{g=0}^{s-1} \sum_{h=0}^{s-1} \sum_{i=0}^{s-1} \sum_{j=0}^{s-1} n_{ghij} \ln(\hat{p}_{ghij}), \tag{8.13d}$$

with obvious extension for fourth-order and higher Markov chains. Here the summations are over all s states of the Markov chain, and so will include only two terms each for two-state (binary) time series. Equation (8.13a) is simply the log-likelihood for the independent binomial model.

Example 8.2. Likelihood Ratio Test for the Order of a Markov Chain

To illustrate the application of Eqs. (8.13), consider a likelihood ratio test of first-order dependence of the time series in Table 8.1, versus the null hypothesis of zero serial correlation. The test involves computation of the log-likelihoods in Eqs. (8.13a) and (8.13b). The resulting two log-likelihoods are compared using the test statistic given by Eq. (5.19).

In the last 30 data points in Table 8.1, there are $n_0 = 14$ zeros and $n_1 = 16$ ones, yielding the unconditional relative frequencies of rain and no rain $\hat{p}_0 = 14/30 = 0.467$ and $\hat{p}_1 = 16/30 = 0.533$, respectively. The last 30 points are used because the first-order Markov chain amounts to two conditional binomial distributions, given the previous day's value, and the value for December 31, 1986 is not available in Table A.1. The log-likelihood in Eq. (8.13a) for these data is $L_0 = 14 \ln(0.467) + 16 \ln(0.533) = -20.73$. Values of n_{ij} and $\hat{p}_{ij}$ were computed previously, and can be substituted into Eq. (8.13b) to yield $L_1 = 11 \ln(0.688) + 5 \ln(0.312) + 4 \ln(0.286) + 10 \ln(0.714) = -18.31$. Necessarily, $L_1 \geq L_0$ because the greater number of parameters in the more elaborate first-order Markov chain model provide more flexibility for a closer fit to the data at hand. The statistical significance of the difference in log-likelihoods can be assessed since it is known that the null distribution of $\Lambda = 2(L_1 - L_0) = 4.83$ is χ^2, with $\nu = [s^{m(H_A)} - s^{m(H_0)}](s - 1)$ degrees of freedom. Since the time series being tested in binary, $s = 2$. The null hypothesis is that the time dependence is zero-order, so

$m(H_0) = 0$, and the alternative hypothesis is first-order serial dependence, or $m(H_A) = 1$. Thus, $\nu = (2^1 - 2^0)(2 - 1) = 1$ degree of freedom. This likelihood test result is consistent with the χ^2 goodness-of-fit test conducted in Example 8.1. ¤

Both the AIC and BIC criteria attempt to find the most appropriate model order by striking a balance between goodness of fit, as reflected in log-likelihoods, and a penalty that increases with the number of fitted parameters. The two approaches differ only in the form of the penalty function. The AIC and BIC statistics are then computed for each trial order m, using

$$\mathrm{AIC}(m) = -2L_m + 2s^m(s - 1), \tag{8.14}$$

or

$$\mathrm{BIC}(m) = -2L_m + s^m(\ln n), \tag{8.15}$$

respectively. The order m is chosen as appropriate that minimizes either Eq. (8.14) or (8.15). The BIC criterion tends to be more conservative, generally picking lower orders than the AIC criterion when results of the two approaches differ. Use of the BIC statistic may be preferable for sufficiently long time series, although "sufficiently long" may range from around $n = 100$ to over $n = 1000$, depending on the nature of the serial correlation (Katz, 1981).

8.3 Time Domain. II. Continuous Data

The Markov chain models described in the previous section are not suitable for describing time series of data that are continuous, in the sense of the data being able to take on infinitely many values on all or part of the real line. As discussed in Chapter 4, atmospheric variables such as temperature, windspeed, and geopotential height are continuous variables. The correlation structure of such time series can often be represented successfully using a class of time-series models known as *Box–Jenkins models*, after the classic text by Box and Jenkins (1976).

8.3.1 First-Order Autoregression

The simplest Box–Jenkins model is the *first-order autoregression* [AR(1)] model. It is the continuous analog of the first-order Markov chain. As the term suggests, one way of viewing the AR(1) model is as a simple linear regression (Chapter 6), where the predictand is the value of the time series at time $t + 1$, x_{t+1}, and the predictor is the current value of the time series, x_t. The AR(1) model can be written as

$$x_{t+1} - \mu = \phi(x_t - \mu) + \epsilon_{t+1}, \tag{8.16}$$

where μ is the mean of the time series, ϕ is the autoregressive parameter, and ϵ_{t+1}

is a random quantity corresponding to the residual in ordinary regression. The right-hand side of Eq. (8.16) consists of a deterministic part in the first term, and a random part in the second term. That is, the next value of the time series x_{t+1} is given by the function of x_t in the first term, plus the random "shock" or "innovation" ϵ_{t+1}.

The time series of x is assumed stationary, so that its mean μ is the same for each interval of time. The data series also exhibits a variance, σ_x^2, the sample counterpart of which is just the ordinary sample variance computed from the values of the time series by squaring Eq. (3.5). The ϵ values are mutually independent random quantities having mean $\mu_\epsilon = 0$ and variance σ_ϵ^2. Very often it is further assumed that the ϵ values follow a Gaussian distribution.

As illustrated in Fig. 8.5, the autoregressive model in Eq. (8.16) can represent the serial correlation of a time series. This is a scatterplot of minimum temperatures at Canandaigua, New York, during January 1987, from Table A.1. Plotted on the horizontal are the first 30 data values, for January 1–30. The corresponding temperatures for the following days, January 2–31, are plotted on the vertical. The expected serial correlation, or persistence, is evident from the appearance of

Fig. 8.5 Scatterplot of January 1–30, 1987 minimum temperatures (°F) at Canandaigua, New York (x_t, horizontal) paired with minimum temperatures for the following days, January 2–31 (x_{t+1}, vertical). The data are from Table A.1. The simple linear regression equation corresponding to the first term of the AR(1) time-series model [Eq. (8.16)] is also shown.

the point cloud, and from the positive slope of the regression line. Equation (8.16) can be viewed as a prediction equation for x_{t+1} using x_t as the predictor. Rearranging Eq. (8.16) to more closely resemble the simple linear regression Eq. (6.3) yields the intercept $a = \mu(1 - \phi)$ and slope $b = \phi$.

Another way to look at Eq. (8.16) is as an algorithm for generating synthetic time series of values of x. Beginning with an initial value, x_0, one would subtract the mean value (i.e., construct the corresponding anomaly), multiply by the autoregressive parameter ϕ, and then add a randomly generated variable ϵ_1 drawn from a Gaussian distribution with mean zero and variance σ_ϵ^2. The first value of the time series, x_1, would then be produced by adding back the mean μ. The next time series value, x_2, would then be produced in a similar way, by operating on x_1 and adding a new random Gaussian quantity ϵ_2. For positive values of the parameter ϕ, synthetic time series constructed in this way will exhibit positive serial correlation because each newly generated data value x_{t+1} includes some information carried forward from the preceding value x_t. Since x_t was in turn generated in part from x_{t-1}, and so on, members of the time series separated by more than one time unit will be correlated, although this correlation becomes progressively weaker as the time separation increases.

The first-order autoregression is sometimes called the *Markov process*, or *Markov scheme*. It shares with the first-order Markov chain the property that the full history of the time series prior to X_t provides no additional information regarding x_{t+1}, once x_t is known. This property can be expressed formally as

$$\begin{aligned}\Pr\{X_{t+1} \le x_{t+1} \mid X_t \le x_t, X_{t-1} \le x_{t-1}, \ldots, X_1 \le x_1\} \\ = \Pr\{X_{t+1} \le x_{t+1} \mid X_t \le x_t\}.\end{aligned} \tag{8.17}$$

Here the notation for continuous random variables has been used to express essentially the same idea as in Eq. (8.1) for a series of discrete events. Again, Eq. (8.17) does not imply that values of the time series separated by more than one time step are independent, but only that the influence of the prior history of the time series on its future values is fully contained in the current value x_t, regardless of the particular path by which the time series arrived at x_t.

Equation (8.16) is also sometimes known as a *red noise* process, because a positive value of the parameter ϕ averages or smooths out short-term fluctuations in the serially independent series of innovations, ϵ, but affects the slower random variations much less strongly. The resulting time series is called red noise by analogy to visible light depleted in the shorter wavelengths, which appears reddish. This topic will be discussed further in Section 8.5, but the effect can be appreciated by looking at Fig. 5.2. This figure compares a series of uncorrelated Gaussian values, ϵ_t (panel a), with an autocorrelated series generated from them using Eq. (8.16) and the value $\phi = 0.6$ (panel b). It is evident that the most erratic point-to-point variations in the uncorrelated series have been smoothed out, while the slower random variations are essentially preserved. In the time domain this

smoothing is expressed as serial correlation. From a frequency perspective, the resulting series is "reddened."

Parameter estimation for the first-order autoregressive model is straightforward. The mean of the time series, μ, is simply the usual sample average [Eq. (1.1)] of the data set, provided the series can be considered to be stationary. Nonstationary series must first be dealt with in one of the ways discussed in Section 8.1.1.

The estimated autoregressive parameter is simply equal to the sample lag-1 autocorrelation coefficient, Eq. (3.23):

$$\hat{\phi} = r_1. \tag{8.18}$$

For the resulting probability model to be stationary, it is required that $-1 < \phi < 1$. As a practical matter this presents no problem for the first-order autoregression, however, because the correlation coefficient is also bounded by the same limits. For most atmospheric time series the parameter ϕ will be positive, reflecting persistence. Negative values of ϕ are possible, but correspond to very jagged (*anticorrelated*) time series with a tendency for alternating values above and below the mean. Because of the Markovian property, the full (theoretical, or "population") autocorrelation function for a time series governed by a first-order autoregressive process can be written in terms of the autoregressive parameter as

$$r_k = \phi^k. \tag{8.19}$$

Thus, the autocorrelation function for an AR(1) process decays exponentially from $r_0 = 1$, approaching zero as $k \to \infty$.

A series of truly independent data would have $\phi = 0$. However, a finite sample of independent data will generally exhibit a nonzero sample estimate of the autoregressive parameter. For a sufficiently long data series the sampling distribution of $\hat{\phi}$ is approximately Gaussian, with $\mu_\phi = \hat{\phi}$ and $\sigma_\phi^2 = (1 - \phi^2)/n$. Therefore, a significance test for the sample estimate of the autoregressive parameter versus the null hypothesis that $\phi = 0$, corresponding to Eq. (5.3), can be carried out using the test statistic

$$z = \frac{\hat{\phi} - 0}{[(1 - \hat{\phi}^2)/n]^{1/2}}, \tag{8.20}$$

with significance assessed approximately using standard Gaussian probabilities. This test is virtually identical to the t test for the slope of a regression line.

The final parameter of the statistical model in Eq. (8.16) is the *residual variance*, or *innovation variance*, σ_ϵ^2. This quantity is sometimes also known as the *white-noise variance*, for reasons to be explained in Section 8.5. This parameter expresses the variability or uncertainty in the time series not accounted for by the serial correlation or, put another way, the uncertainty in x_{t+1} given that x_t is known. The brute-force approach to estimating σ_ϵ^2 is to estimate ϕ using Eq. (8.18), compute the time series ϵ_{t+1} from the data using a rearrangement of

Eq. (8.16), and then to compute the ordinary sample variance of these ϵ values. Since the variance of the data is often computed as a matter of course, another way to estimate the white-noise variance is to use the relationship between the variances of the data series and the innovation series in the AR(1) model:

$$\sigma_\epsilon^2 = (1 - \phi^2)\,\sigma_x^2. \tag{8.21}$$

Equation (8.21) implies $\sigma_\epsilon^2 \leq \sigma_x^2$, with equality only for independent data where $\phi = 0$. Otherwise, Eq. (8.21) implies that knowing the current value of an autocorrelated time series decreases uncertainty about the next value of the time series. In practical settings one works with sample estimates of the autoregressive parameter and of the variance of the data series, so that the corresponding sample estimate of the white-noise variance is

$$s_\epsilon^2 = \frac{1 - \hat{\phi}^2}{n - 2} \sum_{t=1}^{n} (x_t - \bar{x})^2 = \frac{n - 1}{n - 2}(1 - \hat{\phi}^2)s_x^2. \tag{8.22}$$

The difference between Eqs. (8.22) and (8.21) is appreciable only if the data series is relatively short.

Example 8.3. A First-Order Autoregression

Consider fitting an AR(1) process to the series of January 1987 minimum temperatures from Canandaigua, New York, in Table A.1. As indicated in the table, the average of these 31 values is 20.23°F, and this would be adopted as the mean of the time series, assuming stationarity. The sample lag-1 autocorrelation coefficient, from Eq. (3.23), is $r_1 = 0.67$, and this value would be adopted as the estimated autoregressive parameter according to Eq. (8.18). The scatterplot of this data series against itself lagged by one time unit in Fig. 8.5 suggests the positive serial correlation typical of daily temperature data. A formal significance test of the estimated autoregressive parameter versus the null hypothesis that it is really zero would use the test statistic in Eq. (8.20), $z = 0.67/[(1 - 0.67^2)/31]^{1/2} = 5.02$. This test provides extremely strong evidence that the observed sample autocorrelation did not arise only by chance.

The sample standard deviation of the 31 Canandaigua minimum temperatures in Table A.1 is 8.81°F. Using Eq. (8.22), the estimated white-noise variance for the fitted autoregression would be $s_\epsilon^2 = (30/29)(1 - 0.67^2)(8.81^2) = 44.24°\text{F}^2$, corresponding to a standard deviation of 6.65°F. By comparison, the brute-force sample standard deviation of the series of sample residuals, each computed from the rearrangement of Eq. (8.16) as $e_{t+1} = (x_{t+1} - \mu) - \phi(x_t - \mu)$, is 6.55°F.

The computations in this example have been conducted under the assumption that the time series being analyzed is stationary. In particular, it has been assumed that the mean value does not change through time. This assumption is not exactly satisfied by these data, as illustrated in Fig. 8.6. Here the time series of the January 1987 Canandaigua minimum temperature data is shown together with the clima-

Fig. 8.6 Time series of the January 1987 Canandaigua minimum temperature data. Solid line indicates the least-squares linear trend in the data, and the dashed line represents the climatological average minimum temperatures for the period 1961–1990.

tological average temperatures for the period 1961–1990 (dashed line), and the linear trend fit to the 31 data points for 1987 (solid line).

Of course, the dashed line in Fig. 8.6 is a better representation of the long-term (population) mean minimum temperatures at this location, and it indicates that early January is slightly warmer than late January on average. Strictly speaking, then, the data series is not stationary, since the underlying mean value for the time series is not constant through time. However, the change through the month represented by the dashed line is sufficiently minor (in comparison to the variability around this mean function) that one would generally be comfortable in pooling data from a collection of Januarys and assuming stationarity. In fact, the preceding results for the 1987 data sample are not very different if the January 1987 mean minimum temperature of 20.23°F, or the long-term climatological temperatures represented by the dashed line, are assumed. In the latter case, one finds $\phi = 0.64$, and $s_\epsilon^2 = 6.49°\text{F}$.

Because the long-term climatological minimum temperature declines so slowly, it is clear that the rather steep negative slope of the solid line in Fig. 8.6 results mainly from sampling variations in this short example data record. Normally an analysis of this kind would be carried out using a much longer time series. However, if no other information about the January minimum temperature climate of this location were available, it would be sensible to produce a stationary series before proceeding further, by subtracting the mean values represented by the solid line from the data points. The regression equation for this line is $\mu(t) = 29.6 - 0.584t$, where t is the date. The autoregressive process in Eq. (8.16) would then be fit using the time series of the anomalies $x'_t = x_t - \mu(t)$. For example, $x'_1 = 28°\text{F} - (29.6 - 0.584) = -1.02°\text{F}$. Since the average residual from a least-

squares regression line is zero (Chapter 6), the mean of this series of anomalies x'_t will be zero. Fitting Eq. (8.16) to this anomaly series yields $\hat{\phi} = 0.47$, and $s_\epsilon^2 = 39.95°\text{F}^2$. ¤

8.3.2 Higher-Order Autoregressions

The first-order autoregression in Eq. (8.16) generalizes readily to higher orders. That is, the regression equation predicting x_{t+1} can be expanded to include data values progressively further back in time as predictors. The general autoregressive model of order K, or AR(K) model, is

$$x_{t+1} - \mu = \sum_{k=1}^{K} \phi_k (x_{t-k+1} - \mu) + \epsilon_{t+1}. \tag{8.23}$$

Here the anomaly for the next time point, $x_{t+1} - \mu$, is a weighted average of the previous K anomalies plus the random component ϵ_{t+1}, where the weights are the autoregressive coefficients ϕ_k. As before, the ϵ values are mutually independent, with zero mean and variance σ_ϵ^2. Stationarity of the process implies that μ and σ_ϵ^2 do not change through time. For $K = 1$, Eq. (8.23) is identical to Eq. (8.16).

Estimation of the K autoregressive parameters ϕ_k is most easily done using the set of equations relating them to the autocorrelation function, which are known as the *Yule–Walker equations*. These are

$$\left.\begin{array}{lllllll} r_1 = & \hat{\phi}_1 & + \hat{\phi}_2 r_1 & + \hat{\phi}_3 r_2 & + \cdots + & \hat{\phi}_K r_{K-1} \\ r_2 = & \hat{\phi}_1 r_1 & + \hat{\phi}_2 & + \hat{\phi}_3 r_1 & + \cdots + & \hat{\phi}_K r_{K-2} \\ r_3 = & \hat{\phi}_1 r_2 & + \hat{\phi}_2 r_1 & + \hat{\phi}_3 & + \cdots + & \hat{\phi}_K r_{K-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ r_K = & \hat{\phi}_1 r_{K-1} & + \hat{\phi}_2 r_{K-2} & + \hat{\phi}_3 r_{K-3} & + \cdots + & \hat{\phi}_K \end{array}\right\}. \tag{8.24}$$

Here $\hat{\phi}_k = 0$ for $k > K$. The Yule–Walker equations arise from Eq. (8.23), by multiplying by x_t, applying the expected value operator, and evaluating the result for different values of k (e.g., Box and Jenkins, 1976). These equations can be solved simultaneously for the ϕ_k. Alternatively, a method for using these equations recursively for parameter estimation—that is, to compute ϕ_1 and ϕ_2 to fit the AR(2) model knowing ϕ for the AR(1) model, and then to compute ϕ_1, ϕ_2, and ϕ_3 for the AR(3) model knowing ϕ_1 and ϕ_2 for the AR(2) model, and so on—is given in Box and Jenkins (1976) and Katz (1982). Constraints on the autoregressive parameters necessary for Eq. (8.23) to be stationary are given in Box and Jenkins (1976).

The theoretical autocorrelation function corresponding to a particular set of the ϕ_k values can be determined by solving Eq. (8.24) for the first K autocorrelations, and then applying

$$\rho_m = \sum_{k=1}^{K} \phi_k \rho_{m-k}. \tag{8.25}$$

Equation (8.25) holds for lags $m \geq k$, with the understanding that $\rho_0 \equiv 1$. Finally, the generalization of Eq. (8.21) for the relationship between the white-noise variance and the variance of the data values themselves is

$$\sigma_\epsilon^2 = \left(1 - \sum_{k=1}^{K} \phi_k \rho_k\right)\sigma_x^2. \tag{8.26}$$

The AR(2) model A common and important higher-order autoregressive model is the AR(2) process. It is reasonably simple, requiring the fitting of only two parameters in addition to the sample mean and variance of the series, yet it can describe a variety of qualitatively quite different behaviors of time series. The defining equation for AR(2) processes is

$$x_{t+1} - \mu = \phi_1(x_t - \mu) + \phi_2(x_{t-1} - \mu) + \epsilon_{t+1}, \tag{8.27}$$

which is easily seen to be a special case of Eq. (8.23). Using the first $K = 2$ of the Yule–Walker Equations (8.24),

$$r_1 = \hat{\phi}_1 + \hat{\phi}_2 r_1 \tag{8.28a}$$

$$r_2 = \hat{\phi}_1 r_1 + \hat{\phi}_2, \tag{8.28b}$$

one can estimate the two autoregressive parameters as

$$\hat{\phi}_1 = \frac{r_1(1 - r_2)}{1 - r_1^2} \tag{8.29a}$$

and

$$\hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}. \tag{8.29b}$$

Here the estimation equations (8.29) have been obtained simply by solving Eqs. (8.28) for $\hat{\phi}_1$ and $\hat{\phi}_2$.

The white-noise variance for a fitted AR(2) model can be estimated in several ways. For very large samples, Eq. (8.26) with $K = 2$ can be used with the sample variance of the time series, s_ϵ^2. Alternatively, once the autoregressive parameters have been fit using Eqs. (8.29) or some other means, the corresponding estimated time series of the random innovations ϵ can be computed from a rearrangement of Eq. (8.27) and their sample variance computed, as was done in Example 8.3 for the fitted AR(1) process. Another possibility is to use the recursive equation given by Katz (1982),

$$s_\epsilon^2(m) = [1 - \hat{\phi}_m^2(m)]s_\epsilon^2(m - 1). \tag{8.30}$$

Here the autoregressive models AR(1), AR(2), . . . are fitted successively, $s_\epsilon^2(m)$ is the estimated white-noise variance of the mth (i.e., current) autoregression, $s_\epsilon^2(m - 1)$ is the estimated white-noise variance for the previously fitted (one-order-smaller) model, and $\hat{\phi}_m^2(m)$ is the estimated autoregressive parameter for

the highest lag in the current model. For the AR(2) model, Eq. (8.30) can be used with the expression for $s_\epsilon^2(1)$ in Eq. (8.22) to yield

$$s_\epsilon^2(2) = (1 - \hat{\phi}_2^2)\frac{n-1}{n-2}(1 - r_1^2)s_x^2, \tag{8.31}$$

since $\hat{\phi} = r_1$ for the AR(1) model.

For an AR(2) process to be stationary, its two parameters must satisfy the constraints

$$\left.\begin{aligned} \phi_1 + \phi_2 &< 1 \\ \phi_2 - \phi_1 &< 1 \\ -1 < \phi_2 &< 1 \end{aligned}\right\}, \tag{8.32}$$

which define the triangular region in the (ϕ_1, ϕ_2) plane shown in Fig. 8.7. Note that substituting $\phi_2 = 0$ into Eq. (8.32) yields the stationarity condition $-1 < \phi_1 < 1$ applicable to the AR(1) model. Figure 8.7 includes AR(1) models as special cases on the horizontal $\phi_2 = 0$ line, for which this stationarity condition applies.

The first two values of the theoretical autocorrelation function for a particular AR(2) process can be obtained by solving Eqs. (8.28) as

$$\rho_1 = \frac{\phi_1}{1 - \phi_2} \tag{8.33a}$$

Fig. 8.7 The allowable parameter space for stationary AR(2) processes, with insets showing autocorrelation functions for selected AR(2) models. The horizontal $\phi_2 = 0$ line locates the AR(1) models as special cases, and autocorrelation functions for two of these are shown. AR(2) models appropriate to atmospheric time series usually exhibit $\phi_1 > 0$.

and

$$\rho_2 = \phi_2 + \frac{\phi_1^2}{1 - \phi_2}, \tag{8.33b}$$

and subsequent values of the autocorrelation function can be calculated using Eq. (8.25). Figure 8.7 indicates that a wide range of types of autocorrelation functions, and thus a wide range of time correlation behaviors can be represented by AR(2) processes. First, AR(2) models include the simpler AR(1) models as special cases. Two AR(1) autocorrelation functions are shown in Fig. 8.7. The autocorrelation function for the model with $\phi_1 = 0.5$ and $\phi_2 = 0.0$ decays exponentially toward zero, following Eq. (8.19). Autocorrelation functions for many atmospheric time series exhibit this kind of behavior, at least approximately. The other AR(1) model for which an autocorrelation function is shown is for $\phi_1 = -0.6$ and $\phi_2 = 0.0$. Because of the negative lag-1 autocorrelation, the autocorrelation function exhibits oscillations around zero that are progressively damped at longer lags [again, cf. Eq. (8.19)]. That is, there is a tendency for the anomalies of consecutive data values to have opposite signs, so that data separated by even numbers of lags are positively correlated. This kind of behavior is rarely seen in atmospheric data series, and most AR(2) models for atmospheric data have $\phi_1 > 0$.

The second autoregressive parameter allows many other kinds of behaviors in the autocorrelation function. For example, the autocorrelation function for the AR(2) model with $\phi_1 = 0.3$ and $\phi_2 = 0.4$ exhibits a larger correlation at two lags than at one lag. For $\phi_1 = 0.7$ and $\phi_2 = -0.2$ the autocorrelation function declines very quickly, and is almost zero for lags $k \geq 4$. The autocorrelation function for the AR(2) model with $\phi_1 = 0.9$ and $\phi_2 = -0.6$ is very interesting in that it exhibits a slow damped oscillation around zero. This characteristic reflects what are called *pseudoperiodicities* in the corresponding time series. That is, time series values separated by very few lags exhibit fairly strong positive correlation, those separated by a few more lags exhibit negative correlation, whereas values separated by yet a few more lags exhibit positive correlation again. The qualitative effect is for time series to exhibit oscillations around the mean resembling an irregular cosine curve with period approximately equal to the number of lags in the "hump" in the autocorrelation function. Thus, AR(2) models can represent data that are approximately but not strictly periodic, such as barometric pressure variations resulting from the movement of midlatitude synoptic systems.

Some properties of autoregressive models are illustrated by the four example synthetic time series in Fig. 8.8. Series a is simply a sequence of 50 independent Gaussian variates with $\mu = 0$. Series b is a realization of the AR(1) process generated using Eq. (8.16) or, equivalently, Eq. (8.27) with $\mu = 0$, $\phi_1 = 0.5$ and $\phi_2 = 0.0$. The apparent similarity between series a and b arises because series a has been used as the ϵ_{t+1} series forcing the autoregressive process in Eq. (8.27). The effect of the parameter $\phi = \phi_1 > 0$ is to smooth out step-to-step variations in the white-noise series (a), and to give the resulting time series a bit of "memory."

Fig. 8.8 Four synthetic time series illustrating some properties of autoregressive models. Series (a) consists of independent Gaussian variables (white noise). Series (b) is a realization of the AR(1) process with $\phi_1 = 0.5$, and series (c) is a realization of the AR(2) process with $\phi_1 = 0.9$ and $\phi_1 = -0.6$, both of whose autocorrelation functions are shown in Fig. 8.7. Series (d) is nonstationary because its parameters lie outside the triangle in Fig. 8.7, and this nonstationarity can be seen as a drifting in the mean value. Series b–d were constructed using Eq. (8.27) with $\mu = 0$ and the ϵ values from series a.

Series c in Fig. 8.8 is a realization of the AR(2) process with $\mu = 0$, $\phi_1 = 0.9$, and $\phi_2 = -0.6$. It resembles qualitatively some atmospheric series (e.g., mid-latitude sea-level pressures), but has been generated using Eq. (8.27) with series a as the forcing white noise. This series exhibits pseudoperiodicities. That is, peaks and troughs in this time series tend to recur with a period near six or seven time intervals, but these are not so regular that a cosine function or the sum of a few cosine functions would represent them very well. This feature is the expression in the data series of the hump in the autocorrelation function for this autoregressive model shown in the inset of Fig. 8.7, which occurs at a lag interval of six or seven time periods. Similarly, the peak/trough pairs tend to be separated by perhaps three or four time intervals, corresponding to the minimum in the autocorrelation function at these lags shown in the inset in Fig. 8.7.

The autoregressive parameters $\phi_1 = 0.9$ and $\phi_2 = 0.11$ for series d in Fig. 8.8 fall outside the triangular region in Fig. 8.7 that define the limits of stationary

AR(2) processes. The series is therefore not stationary, and the nonstationarity can be seen as a drifting of the mean value in the realization of this process shown in Fig. 8.8.

Finally, series a–c in Fig. 8.8 illustrate the nature of the relationship between the variance of the time series, σ_x^2, and the white-noise variance, σ_ϵ^2, of an autoregressive process. Series a consists simply of independent Gaussian variates, or white noise. Formally, it can be viewed as a special case of an autoregressive process, with all the ϕ_k values equal to zero. Using Eq. (8.26), it is clear that $\sigma_x^2 = \sigma_\epsilon^2$ for this series. Since series b and c were generated using series a as the white noise forcing ϵ_{t+1}, σ_ϵ^2 for all three of these series are equal. Time series c gives the visual impression of being more variable than series b, which in turn appears to exhibit more variability than series a. Using Eqs. (8.33) with Eq. (8.26), it is easy to compute that σ_x^2 for series b is 1.33 times larger than the common σ_ϵ^2, and that for series c is 2.29 times larger. The equations on which these computations are based pertain only to stationary autoregressive series, and so cannot be meaningfully applied to the nonstationary series d.

8.3.3 Order Selection Criteria

The Yule–Walker equations, Eqs. (8.24), can be used to fit autoregressive models to essentially arbitrarily high order. At some point, however, expanding the complexity of the model will not appreciably improve its representation of the data. Arbitrarily adding more terms in Eq. (8.23) will eventually result in the model being overfit, or excessively "tuned" to the data used for parameter estimation.

The BIC (Schwarz, 1978) and AIC (Akaike, 1974) statistics, applied to Markov chains in Section 8.2, are also often used to decide among potential orders of autoregressive models. Both statistics involve the log-likelihood plus a penalty for the number of parameters, with the two criteria differing only in the form of the penalty function. Here the likelihood function involves the estimated white-noise variance.

For each candidate order m, the order selection statistics

$$\mathrm{BIC}(m) = n \ln\left[\frac{n}{n - m - 1} s_\epsilon^2(m)\right] + (m + 1) \ln n, \tag{8.34}$$

or

$$\mathrm{AIC}(m) = n \ln\left[\frac{n}{n - m - 1} s_\epsilon^2(m)\right] + 2(m + 1), \tag{8.35}$$

are computed, using $s_\epsilon^2(m)$ from Eq. (8.30). Better-fitting models will exhibit smaller white-noise variance, implying less residual uncertainty. Arbitrarily adding more parameters (fitting higher- and higher-order autoregressive models) will not increase the white-noise variance estimated from the data sample, but neither

will the estimated white-noise variance decrease much if the extra parameters are not effective in describing the behavior of the data. Thus, the penalty functions serve to guard against overfitting. That order m is chosen as appropriate that minimizes either Eq. (8.34) or (8.35).

Example 8.4. Order Selection among Autoregressive Models

Table 8.3 summarizes the results of fitting successively higher autoregressive models to the January 1987 Canandaigua minimum temperature data, assuming that they are stationary without removal of a trend. The first column shows the sample autocorrelation function up to lag 7. The estimated white-noise variance for autoregressions of orders 1–7, computed using the Yule–Walker equations and Eq. (8.30), are shown in the third column. Notice that $s_\epsilon^2(0)$ is simply the sample variance of the time series itself, or s_x^2. The estimated white-noise variances decrease progressively as more terms are added to Eq. (8.23), but toward the bottom of the table adding yet more terms has little further effect.

The BIC and AIC statistics for each candidate autoregression are shown in the last two columns. Both indicate that the AR(1) model is most appropriate for these data, as this produces the minimum in both order selection statistics. Similar results also are obtained for the other three temperature series in Table A.1. Note, however, that with a larger sample size, higher-order autoregressions could be chosen by both criteria. For the estimated residual variances shown in Table 8.3, using the AIC statistic would lead to the choice of the AR(2) model for n greater than about 290, while the AR(2) model would minimize the BIC statistic for n larger than about 430. ¤

Table 8.3

Order Selection for Autoregressive Models to Represent the January 1987 Canandaigua Minimum Temperature Series, Assuming Stationarity[a]

Lag, m	r_m	$s_\epsilon^2(m)$	BIC (m)	AIC (m)
0	1.000	77.58	138.32	136.89
1	0.672	42.55	125.20	122.34
2	0.507	42.11	129.41	125.11
3	0.397	42.04	133.91	128.18
4	0.432	39.72	136.76	129.59
5	0.198	34.39	136.94	128.34
6	0.183	33.03	140.39	130.35
7	0.161	33.02	145.14	133.66

[a] Presented are the autocorrelation function for the first seven lags m, the estimated white-noise variance for each AR(m) model, and the BIC and AIC statistics for each trial order. For $m = 0$ the autocorrelation function is 1.00, and the white-noise variance is equal to the sample variance of the series. The AR(1) model is selected by both the BIC and AIC criteria.

8.3.4 The Variance of a Time Average

An important application of time-series models in atmospheric data analysis is estimation of the sampling distribution of the average of a correlated time series. Recall that a sampling distribution characterizes the batch-to-batch variability of a statistic computed from finite data samples. If the data values making up a sample average are independent, the variance of the sampling distribution of that average is given by the variance of the data, s_x^2, divided by the sample size [Eq. (5.4)].

Since atmospheric data are often positively correlated, using Eq. (5.4) to calculate the variance of (the sampling distribution of) a time average leads to an underestimate. This is a consequence of the tendency for nearby values of correlated time series to be similar, leading to less batch-to-batch consistency of the sample average. As discussed in Chapter 5, the resulting underestimates of the variance of the sampling distribution of the mean can lead to serious problems for hypothesis tests, leading in general to unwarranted rejection of null hypotheses.

The effect of serial correlation on the variance of a time average over a sufficiently large sample can be accounted for through a variance inflation factor, V, modifying Eq. (5.4):

$$\mathrm{Var}[\bar{x}] = \frac{V\sigma_x^2}{n}. \tag{8.36}$$

If the data series is uncorrelated, $V = 1$ and Eq. (8.36) corresponds to Eq. (5.4). If the data exhibit positive serial correlation, $V > 1$ and the variance of the time average is inflated above what would result from independent data. Note, however, that even if the underlying data are correlated, the mean of the sampling distribution of the time average is the same as the underlying mean of the data being averaged:

$$E[\bar{x}] = \mu_{\bar{x}} = E[x_t] = \mu_x. \tag{8.37}$$

For large sample size, the variance inflation factor depends on the autocorrelation function according to

$$V = 1 + 2\sum_{k=1}^{\infty} \rho_k. \tag{8.38}$$

However, the variance inflation factor can be estimated with greater ease and precision if the data series is well represented by an autoregressive model. In terms of the parameters of an AR(K) model, the large-sample variance inflation factor in Eq. (8.38) is

$$V = \frac{1 - \sum_{k=1}^{K} \phi_k \rho_k}{[1 - \sum_{k=1}^{K} \phi_k]^2}. \tag{8.39}$$

Note that the theoretical autocorrelations ρ_k in Eq. (8.39) can be expressed in terms of the autoregressive parameters by solving the Yule–Walker equations (8.24) for the correlations. In the special case of an AR(1) model being appropriate for a time series of interest, Eq. (8.39) reduces to

$$V = \frac{1 + \phi_1}{1 - \phi_1}, \tag{8.40}$$

which was used to estimate the "effective sample size" in Eq. (5.12), and the variance of the sampling distribution of a sample mean in Eq. (5.13). Equation (8.40) is a convenient large-sample approximation to the formula for the AR(1) variance inflation factor

$$V = \frac{1 + \phi_1 - 2\phi_1^n}{1 - \phi_1}, \tag{8.41}$$

which approaches Eq. (8.40) for large sample size n, but yields $V = 1$ for $n = 1$. More generally, if a higher-order autoregressive model is more appropriate to a time series of interest, Eq. (8.39) should be used to compute the variance inflation factor.

Example 8.5. Variances of Time Averages of Different Lengths

The relationship between the variance of a time average and the variance of the individual elements of a time series in Eq. (8.36) can be useful in surprising ways. Consider, for example, the average winter (December–February) geopotential heights, and the standard deviation of the average for the Northern Hemisphere shown in Figs. 8.9a and 8.9b, respectively. Figure 8.9a shows the average field [Eq. (8.37)], and Fig. 8.9b shows the standard deviation of 90-day averages of winter 500-mb heights, representing the interannual variability. That is, Fig. 8.9b shows the square root of Eq. (8.36), with s_x^2 being the variance of the daily 500-mb height measurements and $n = 90$.

Suppose, however, that the sampling distribution of 500-mb heights averaged over a different length of time were needed. One might be interested in the variance of 10-day averages of 500-mb heights at selected locations, for use as a climatological reference for calculating the skill of forecasts of 10-day averages of this quantity, using Eq. (7.38). (Note that the variance of the climatological distribution is exactly the mean-squared error of the climatological reference forecast.) Assuming that time series of winter 500-mb heights are stationary, one can approximate the variance of an average over some different time period without explicitly knowing the variance inflation factor in either Eq. (8.38) or (8.39), and therefore without necessarily having to know the daily data. The ratio of the variances of 10-day and 90-day averages can be constructed from Eq. (8.36),

$$\frac{\mathrm{Var}[\bar{x}_{10}]}{\mathrm{Var}[\bar{x}_{90}]} = \frac{Vs_x^2/10}{Vs_x^2/90}, \tag{8.42a}$$

Fig. 8.9 Average 500-mb-height field for the Northern Hemisphere winter (a) and the field of standard deviations of that average, reflecting winter-to-winter variations (b). [From Blackmon (1976), reproduced with permission of the American Meteorological Society.]

leading to

$$\mathrm{Var}[\bar{x}_{10}] = \frac{90}{10}\,\mathrm{Var}[\bar{x}_{90}]. \tag{8.42b}$$

Regardless of the averaging period, the variance inflation factor V and the variance of the daily observations s_x^2 are the same because they are characteristics of the underlying daily time series. Thus, the variance of a 10-day average is approximately nine times larger than the variance of a 90-day average, and a map of hemispheric 10-day standard deviations of winter 500-mb heights would be qualitatively very similar to Fig. 8.9b, but exhibiting magnitudes about $\sqrt{9} = 3$ times larger. For short averaging periods (n), more accurate estimates could result from using the exact variance inflation factor for an AR(1) process [Eq. (8.41)], rather than assuming equality of V for the two averaging periods as done in Eq. (8.42). ¤

8.3.5 Autoregressive Moving-Average Models

Autoregressive models actually constitute a subset of a broader class of time-domain models, known as *autoregressive moving-average* (ARMA) models. The general ARMA(K,M) model has K autoregressive terms, as in the AR(K) Eq. (8.23), and in addition contains M moving-average terms that constitute a weighted average of the M previous ϵ values. The ARMA(K,M) model thus contains K autoregressive parameters ϕ_k and M moving-average parameters θ_m that affect the time series according to

$$x_{t+1} - \mu = \sum_{k=1}^{K} \phi_k(x_{t-k+1} - \mu) + \epsilon_{t+1} + \sum_{m=1}^{M} \theta_m \epsilon_{t-m+1}. \tag{8.43}$$

The AR(K) process in Eq. (8.23) is a special case of the ARMA(K,M) model in Eq. (8.43), with all the $\theta_m = 0$. Similarly, a pure *moving-average process* of order M, or MA(M) process, would be a special case of Eq. (8.43), with all the $\phi_k = 0$. Parameter estimation and derivation of the autocorrelation function for the general ARMA(K,M) process is more difficult than for the simpler AR(K) models. Parameter estimation methods are given in Box and Jenkins (1976), and many time-series computer packages will fit ARMA models.

An important and common ARMA model is the ARMA(1,1) process,

$$x_{t+1} - \mu = \phi_1(x_t - \mu) + \epsilon_{t+1} + \theta_1\epsilon_t. \tag{8.44}$$

Computing parameter estimates even for this simple ARMA model is somewhat complicated, although Box and Jenkins (1976) present an easy graphical technique that allows estimation of ϕ_1 and θ_1 using the first two sample lag correlations r_1 and r_2. The autocorrelation function for the ARMA (1,1) process can be calculated from the parameters using

$$\rho_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\phi_1\theta_1} \tag{8.45a}$$

and

$$\rho_k = \phi_1 \rho_{k-1}, \qquad k > 1. \tag{8.45b}$$

The autocorrelation function of an ARMA(1,1) process decays exponentially from its value at ρ_1, which depends on both ϕ_1 and θ_1. This differs from the autocorrelation function for an AR(1) process, which decays exponentially from $\rho_1 = 1$. The relationship between the time-series variance and the white-noise variance of an ARMA(1,1) process is

$$\sigma_\epsilon^2 = \frac{1 - \phi_1^2}{1 + \theta_1^2 + 2\phi_1\theta_1}\sigma_x^2. \tag{8.46}$$

Equations (8.45) and (8.46) also apply to the simpler AR(1) and MA(1) processes, for which $\theta_1 = 0$ or $\phi_1 = 0$, respectively.

8.3.6 Simulation and Forecasting with Time-Domain Models

An important application of time-domain models is in the simulation of synthetic (i.e., random-number) series having statistical characteristics that are similar to observed atmospheric data. These Monte Carlo simulations are useful for investigating impacts of atmospheric variability in situations where the record length of the observed data are known or suspected to be insufficient to include representative sequences of the relevant atmospheric variable(s). Here it is necessary to choose carefully the type and order of time-series model, so that the simulated time series will represent the variability of the real atmosphere well.

Once an appropriate time-series model has been identified and its parameters estimated, its defining equation can be used as an algorithm to generate synthetic time series. For example, if an AR(2) model is representative of the data, Eq. (8.27) would be used, whereas Eq. (8.44) would be used as the generation algorithm for ARMA(1,1) models. The simulation method is similar to that described above for sequences of binary variables generated using the Markov chain model. Here, however, the "noise" or "innovation" series, ϵ_{t+1}, is usually assumed to consist of independent Gaussian variates with $\mu_\epsilon = 0$ and variance σ_ϵ^2, which is estimated from the data as described above.

At each time step, a new Gaussian ϵ_{t+1} is chosen and substituted into the defining equation. The next value of the synthetic time series x_{t+1} is then computed using the previous K values of x (for AR models), the previous M values of ϵ (for MA models), or both (for ARMA models). The only real difficulty in implementing the process is at the beginning of each synthetic series, where there are no prior values of x and/or ϵ that can be used. A simple solution to this problem is to substitute the corresponding averages (expected values) for the unknown previous values. That is, one assumes $(x_t - \mu) = 0$ and $\epsilon_t = 0$ for $t \leq 0$.

A better procedure is to generate the first values in a way that is consistent with the structure of the time-series model. For example, with an AR(1) model one could choose x_0 from a Gaussian distribution with variance $\sigma_x^2 = \sigma_\epsilon^2/(1 - \phi^2)$

[cf. Eq. (8.21)]. Another very workable solution is to begin with $(x_t - \mu) = 0$ and $\epsilon_t = 0$, but generate a longer time series than needed. The first few members of the resulting time series, which are most influenced by the initial values, are then discarded.

Example 8.6. Statistical Simulation with an Autoregressive Model

The time series in Figs. 8.8b–8.8d were produced according to the procedure just described, using the independent Gaussian series in Fig. 8.8a as the series of ϵ values. The first and last few values of this independent series, and of the two series plotted in Figs. 8.8b and 8.8c, are given in Table 8.4. For all three series, $\mu = 0$ and $\sigma_\epsilon^2 = 1$. Equation (8.16) has been used to compute the values of the AR(1) series, with $\phi_1 = 0.5$, and Eq. (8.27) was used to compute the AR(2) series, with $\phi_1 = 0.9$ and $\phi_1 = -0.6$.

Consider the more difficult case of generating the AR(2) series. Calculating x_1 and x_2 in order to begin the series presents an immediate problem, because "x_0" and "x_{-1}" do not exist. This simulation was initialized by assuming the expected values $E[x_0] = E[x_{-1}] = \mu = 0$. Thus, since $\mu = 0$, it follows that $x_1 = \phi_1 x_0 + \phi_2 x_{-1} + \epsilon_1 = (0.9)(0) - (0.6)(0) + 1.526 = 1.562$. Having been computed in this way, x_1 is then used to obtain $x_2 = \phi_1 x_1 + \phi_2 x_0 + \epsilon_2 = (0.9)(1.526) - (0.6)(0) + 0.623 = 1.996$. For values of the AR(2) series at times $t = 3$ and larger, the computation is a straightforward application of Eq. (8.27). For example, $x_3 = \phi_1 x_2 + \phi_2 x_1 + \epsilon_3 = (0.9)(1.996) - (0.6)(1.526) - 0.272 = 0.609$. Similarly, $x_4 = \phi_1 x_3 + \phi_2 x_2 + \epsilon_4 = (0.9)(0.609) - (0.6)(1.996) + 0.092 = -0.558$. ¤

Purely statistical forecasts of the future evolution of time series can be produced using time-domain models. These are accomplished by simply extrapolating the

Table 8.4

Values of the Time Series Plotted in Fig. 8.8a–8.8c[a]

t	Independent Gaussian series, ε_t (Fig. 8.8a)	AR(1) series, x_t (Fig. 8.8b)	AR(2) series, x_t (Fig. 8.8c)
1	1.526	1.526	1.526
2	0.623	1.387	1.996
3	−0.272	0.421	0.609
4	0.092	0.302	−0.558
5	0.823	0.974	−0.045
⋮	⋮	⋮	⋮
49	−0.505	−1.073	−3.172
50	−0.927	−1.463	−2.648

[a] The AR(1) and AR(2) series have been generated from the independent Gaussian series using Eqs. (8.16) and (8.27), respectively, as the algorithm.

most recently observed value(s) into the future using the defining equation for the appropriate model, on the basis of parameter estimates fitted from the previous history of the series. Since the future values of ϵ cannot be known, the extrapolations are made using their expected values: $E[\epsilon] = 0$.

The nature of this kind of forecast is most easily appreciated for the AR(1) model, the defining equation for which is Eq. (8.16). Assume that the mean μ and the autoregressive parameter ϕ have been estimated from a time series of observations, the most recent of which is x_t. The forecast for x_{t+1} would be made by setting the unknown future ϵ_{t+1} to zero, and rearranging Eq. (8.16) to yield $x_{t+1} = \mu + \phi(x_t - \mu)$. Note that, in common with the forecasting of a binary time series using a Markov chain model, this forecast is a compromise between persistence ($x_{t+1} = x_t$, which would result if $\phi = 1$) and climatology ($x_{t+1} = \mu$, which would result if $\phi = 0$). Further projections into the future would be obtained by extrapolating the previously forecast values, for example, $x_{t+2} = \mu + \phi(x_{t+1} - \mu)$, and $x_{t+3} = \mu + \phi(x_{t+2} - \mu)$. For the AR(1) model and $\phi > 0$, this series of forecasts would exponentially approach $x_\infty = \mu$.

The same procedure is used for higher-order autoregressions, except that the most recent K values of the time series are needed to extrapolate an AR(K) process [Eq. (8.23)]. Forecasts derived from an AR(2) model, for example, would be made using the previous two observations of the time series, or $x_{t+1} = \mu + \phi_1(x_t - \mu) + \phi_2(x_{t-1} - \mu)$. Forecasts using ARMA models are only slightly more difficult, requiring that the last M values of the ϵ series be back-calculated before the projections begin.

Forecasts made using time-series models are, of course, uncertain, and the forecast uncertainty increases for longer projections into the future. This uncertainty also depends on the nature of the appropriate time-series model (e.g., the order of an autoregression and its parameter values) and on the intrinsic uncertainty in the random-noise series that is quantified by the white-noise variance σ_ϵ^2. The variance of a forecast made only one time step into the future is simply equal to the white-noise variance. Assuming that the ϵ values follow a Gaussian distribution, a 95% confidence interval on a forecast of x_{t+1} is then approximately $x_{t+1} \pm 2\sigma_\epsilon$. For very long extrapolations, the variance of the forecasts approaches the variance of the time series itself, σ_x^2, which for AR models can be computed from the white-noise variance using Eq. (8.26).

For intermediate time projections, calculation of forecast uncertainty is more complicated. For a forecast j time units into the future, the variance of the forecast is given by

$$\sigma^2(x_{t+j}) = \sigma_\epsilon^2\left[1 + \sum_{i=1}^{j-1} \psi_i^2\right]. \tag{8.47}$$

Here the weights ψ_i depend on the parameters of the time-series model, so that Eq. (8.47) indicates the variance of the forecast increases with both the white-

noise variance and the projection, and the increase in uncertainty at increasing lead time depends on the nature of the particular time-series model. For the $j = 1$ time-step forecast, there are no terms in the summation in Eq. (8.47), and the forecast variance is equal to the white-noise variance, as noted above.

For AR(1) models, the ψ weights are simply

$$\psi_i = \phi^i, \qquad i > 0, \tag{8.48}$$

so that, for example, $\psi_1 = \phi$, $\psi_2 = \phi^2$, and so on. More generally, for AR(K) models, the ψ weights are computed recursively, using

$$\psi_i = \sum_{k=1}^{K} \phi_k \psi_{i-k}, \tag{8.49}$$

where it is understood that $\psi_0 = 1$ and $\psi_i = 0$ for $i < 0$. For AR(2) models, for example, $\psi_1 = \phi_1$, $\psi_2 = \phi_1^2 + \phi_2$, $\psi_3 = \phi_1(\phi_1^2 + \phi_2) + \phi_2\phi_1$, and so on. Equations that can be used to compute the ψ weights for MA and ARMA models are given in Box and Jenkins (1976).

Example 8.7. Forecasting with an Autoregressive Model

Figure 8.10 illustrates forecasts using the AR(2) model with $\phi_1 = 0.9$ and $\phi_2 = -0.6$. The first six points in the time series, shown by the circles connected by heavy lines, are the same as the final six points in the time series shown in Fig. 8.8c. The extrapolation of this series into the future, using Eq. (8.27) with all $\epsilon_{t+1} = 0$, is shown by the continuation of the heavy line connecting the crosses. Note that only the final two "observed" values are used to extrapolate the series. The forecast series continues to show the pseudoperiodic behavior characteristic of this AR(2) model, but its oscillations damp out at longer projections as the forecast series approaches the mean μ.

The approximate 95% confidence intervals for the forecast time-series values, given by $\pm 2\sigma(X_{t+j})$ as computed from Eq. (8.47), are shown by the dashed lines. For the particular values of the autoregressive parameters $\phi_1 = 0.9$ and $\phi_2 = -0.6$, Eq. (8.49) yields $\psi_1 = 0.90$, $\psi_2 = 0.21$, $\psi_3 = -0.35$, $\psi_4 = -0.44$, and so on. Note that the confidence band follows the oscillations of the forecast series, and broadens from $\pm 2\sigma_\epsilon$ at a projection of one time unit to nearly $\pm 2\sigma_x$ at longer projections.

Finally, Fig. 8.10 shows the relationship between the forecast time series, and the first five points of three realizations of this AR(2) process, shown by the thin lines connecting points labeled 1, 2, and 3. Each of these three series was computed using Eq. (8.27), starting from $x_t = -2.648$ and $x_{t-1} = -3.172$, but using different sequences of independent Gaussian ϵ values. For the first two or three projections these remain reasonably close to the forecasts. Subsequently the three series begin to diverge as the influence of the final two points from Fig. 8.8c diminishes and the accumulated influence of the new (and different) random ϵ

Fig. 8.10 The final six points of the AR(2) time series in Fig. 8.8c (heavy line, with circles), and its forecast evolution (heavy line, with crosses) extrapolated using Eq. (8.27) with all the $\epsilon = 0$. The $\pm 2\sigma$ limits describing the uncertainty of the forecast values are shown with dashed lines. These standard deviations depend on the forecast lead time. For the one-step-ahead forecast, the width of the confidence interval is a function simply on the white-noise variance, $\pm 2\sigma_\epsilon$. For very long lead times, the forecasts converge to the mean, μ, of the process, and the width of the confidence interval increases to $\pm 2\sigma_x$. Three example realizations of the first five points of the future evolution of the time series, computed using Eq. 8.27 and particular random ϵ values, are also shown (thin lines connecting numbered points).

values increases. For clarity these series have not been plotted more than five time units into the future, although doing so would have shown each to oscillate irregularly, with progressively less relationship to the forecast series. ¤

8.3.7 Multivariate Autoregressions

The AR(K) processes pertaining to scalar time series generalize to the representation of the time dependence exhibited by multivariate series. In this case the

variable $\mathbf{x}$ is a vector quantity, and the generalization of Eq. (8.23) in matrix algebra notation (see Section 9.2) is

$$\mathbf{x}_{t+1} - \boldsymbol{\mu} = \sum_{k=1}^{K} [\boldsymbol{\Phi}_k](\mathbf{x}_{t-k+1} - \boldsymbol{\mu}) + [\mathbf{B}]\boldsymbol{\epsilon}_{t+1}. \tag{8.50}$$

Here the elements of the vector $\mathbf{x}$ consist of a set of correlated time series, the vector $\boldsymbol{\mu}$ consists of the corresponding mean values of the elements of $\mathbf{x}$, and the elements of $\boldsymbol{\epsilon}$ are mutually independent (and usually Gaussian) random variables with unit variance. The matrices of autoregressive parameters $[\boldsymbol{\Phi}_k]$ correspond to the scalar autoregressive parameters ϕ_k in Eq. (8.23). The matrix $[\mathbf{B}]$, operating on the vector $\boldsymbol{\epsilon}_{t+1}$, allows the random components in Eq. (8.50) to have different variances, and to be mutually correlated at each time step (although uncorrelated in time).

Most often used is the multivariate AR(1) process,

$$\mathbf{x}_{t+1} - \boldsymbol{\mu} = [\boldsymbol{\Phi}](\mathbf{x}_t - \boldsymbol{\mu}) + [\mathbf{B}]\boldsymbol{\epsilon}_{t+1}, \tag{8.51}$$

which is the multivariate generalization of Eq. (8.16). Bras and Rodríguez-Iturbe (1985) describe use of Eq. (8.51) in hydrology, where the elements of $\mathbf{x}$ are simultaneously measured streamflows at different locations. This equation is also often used as part of a "synthetic weather generator" (Richardson, 1981). In this second application $\mathbf{x}$ has three elements, which are daily maximum temperature, minimum temperature, and solar radiation for a given location.

The two parameter matrices in Eq. (8.51) are most easily estimated using the simultaneous and lagged correlations among the elements of $\mathbf{x}$. The simultaneous correlations are contained in the correlation matrix $[\mathbf{R}]$ [Fig. 3.17 and Eq. (9.21)], and the lagged correlations are contained in the matrix

$$[\mathbf{R}_1] = \frac{1}{n-1}[\mathbf{Z}_t]^{\mathrm{T}}[\mathbf{Z}_{t+1}] \tag{8.52a}$$

$$= \begin{bmatrix} r_1(1 \to 1) & r_1(2 \to 1) & \cdots & r_1(I \to 1) \\ r_1(1 \to 2) & r_1(2 \to 2) & \cdots & r_1(I \to 2) \\ \vdots & \vdots & & \vdots \\ r_1(1 \to I) & r_1(2 \to I) & \cdots & r_1(I \to I) \end{bmatrix}. \tag{8.52b}$$

The matrices $[\mathbf{Z}]$ in Eq. (8.52a) are defined in the same way as those in Eq. (9.21), although the elements of the standardized data matrices $[\mathbf{Z}_{t+1}]$ and $[\mathbf{Z}_t]$ are shifted relative to each other by one time interval. The diagonal elements of $[\mathbf{R}_1]$ shown in Eq. (8.52b) are the lag-1 autocorrelations [Eq. (3.23)] for each of the I elements of $\mathbf{x}$. The off-diagonal elements of $[\mathbf{R}_1]$ are the correlations among the elements of $\mathbf{x}$ when shifted by one time interval. The arrow notation in this equation indicates the time sequence of the lagging of the variables. For example, $r_1(1 \to 2)$ denotes the correlation between x_1 at time t, and x_2 at time $t + 1$,

whereas r_1 (2→1) denotes the correlation between x_2 at time t, and x_1 at time $t+1$. Notice that the matrix [**R**] is symmetrical, but that in general [$\mathbf{R}_1$] is not.

The matrix of autoregressive parameters [Φ] in Eq. (8.51) is obtained from the two correlation matrices using

$$[\mathbf{\Phi}] = [\mathbf{R}_1][\mathbf{R}]^{-1}. \tag{8.53}$$

Obtaining the matrix [**B**] is somewhat more difficult, although any matrix satisfying the relationship

$$[\mathbf{B}][\mathbf{B}]^{\mathrm{T}} = [\mathbf{R}] - [\mathbf{\Phi}][\mathbf{R}_1]^{\mathrm{T}} \tag{8.54}$$

can be used. Usually the easiest approach to dealing with Eq. (8.54) is the *Cholesky decomposition* (or *square-root* matrix) of [**B**][**B**]^{T} (e.g., Atkinson, 1978; Bras and Rodríguez-Iturbe, 1985), which yields a solution [**B**] whose elements above the diagonal are all zero.

8.4 Frequency Domain. I. Harmonic Analysis

Analysis in the frequency domain involves representing data series in terms of contributions made at different time scales. For example, a time series of hourly temperature data from a midlatitude location will usually exhibit strong variations both at the daily time scale (corresponding to the diurnal cycle of solar heating) and at the annual time scale (reflecting the march of the seasons). In the time domain, these cycles would appear as large positive values in the autocorrelation function for lags at and near 24 h for the diurnal cycle, and $24 \times 365 = 8760$ h for the annual cycle. Thinking about the same time series in the frequency domain, one speaks of large contributions to the total variability of the time series at periods of 24 and 8760 h, or at frequencies of $1/24 = 0.0417\ \mathrm{h}^{-1}$ and $1/8760 = 0.000114\ \mathrm{h}^{-1}$.

Harmonic analysis consists of representing the fluctuations or variations in a time series as having arisen from the adding together of a series of sine and cosine functions. These trigonometric functions are "harmonic" in the sense that they are chosen to have frequencies exhibiting integer multiples of the "fundamental" frequency determined by the sample size of the data series. A common physical analogy is the musical sound produced by a vibrating string, where the pitch is determined by the fundamental frequency, but the aesthetic quality of the sound depends also on the relative contributions of the higher harmonics.

8.4.1 Cosine and Sine Functions

It is worthwhile to review briefly the nature of the cosine function $\cos(\alpha)$ and the sine function $\sin(\alpha)$. The argument in both is a quantity α, measured in angular units. These units can be either degrees or radians. Figure 8.11 shows portions of

Fig. 8.11 Portions of the cosine (solid) and sine (dashed) functions on the interval 0°–450° or, equivalently, 0 to $5\pi/2$ radians. Each executes a full cycle every 360°, or 2π radians, and extends left to $-\infty$ and right to $+\infty$.

the cosine (solid) and sine (dashed) functions, on the angular interval 0 to $5\pi/2$ radians (0°–450°).

The cosine and sine functions extend through indefinitely large negative and positive angles. The same wave pattern repeats every 2π radians or 360°, so that

$$\cos(2\pi k + \alpha) = \cos(\alpha), \tag{8.55}$$

where k is any integer. An analogous equation holds for the sine function. That is, both cosine and sine functions are periodic. Both functions oscillate around their average value of zero, and attain maximum values of $+1$ and minimum values of -1. The cosine function is maximized at 0°, 360°, and so forth, and the sine function is maximized at 90°, 450°, and so on.

These two functions have exactly the same shape but are offset from each other by 90°. Sliding the cosine function to the right by 90° produces the sine function, and sliding the sine function to the left by 90° produces the cosine function:

$$\cos\left(\alpha - \frac{\pi}{2}\right) = \sin(\alpha) \tag{8.56a}$$

and

$$\sin\left(\alpha + \frac{\pi}{2}\right) = \cos(\alpha). \tag{8.56b}$$

8.4.2 Representing a Simple Time Series with a Harmonic Function

Even in the simple situation of time series having a sinusoidal character and executing a single cycle over the course of n observations, three small difficulties must be overcome in order to use a sine or cosine function to represent it:

1. The argument of a trigonometric function is an angle, whereas the data series is a function of time.

2. Cosine and sine functions fluctuate between $+1$ and -1, whereas the data will generally fluctuate between different limits.
3. The cosine function is at its maximum value for $\alpha = 0$ and $\alpha = 2\pi$, and the sine function is at its mean value for $\alpha = 0$ and $\alpha = 2\pi$. Both the sine and cosine may thus be positioned arbitrarily in the horizontal with respect to the data.

The solution to the first problem comes through regarding the length of the data record, n, as constituting a full cycle, or the *fundamental period.* Since the full cycle corresponds to 360° or 2π radians in angular measure, it is easy to proportionally rescale time to angular measure, using

$$\alpha = \left(\frac{360°}{\text{cycle}}\right)\left(\frac{t \text{ time units}}{n \text{ time units/cycle}}\right) = \frac{t}{n}\,360° \tag{8.57a}$$

or

$$\alpha = \left(\frac{2\pi}{\text{cycle}}\right)\left(\frac{t \text{ time units}}{n \text{ time units/cycle}}\right) = 2\pi\,\frac{t}{n}. \tag{8.57b}$$

These equations can be viewed as specifying the angle that subtends proportionally the same part of the distance between 0 and 2π, as the point t is located in time between 0 and n. The quantity

$$\omega_1 = \frac{2\pi}{n} \tag{8.58}$$

is called the *fundamental frequency.* This quantity is an angular frequency and has physical dimensions of radians per unit time. The fundamental frequency specifies the fraction of the full cycle, spanning n time units, that is executed during a single time unit. The subscript "1" indicates that ω_1 pertains to the wave that executes one full cycle over the whole data series.

The second problem is overcome by shifting the cosine or sine function up or down to the general level of the data, and then stretching or compressing it vertically until its range corresponds to that of the data. Since the mean of a pure cosine or sine wave is zero, simply adding the mean value of the data series to the cosine function assures that it will fluctuate around that mean value. The stretching or shrinking is accomplished by multiplying the cosine function by a constant, C_1, known as the *amplitude.* Again, the subscript "1" indicates that this is the amplitude of the fundamental harmonic. Since the maximum and minimum values of a cosine function are ± 1, the maximum and minimum values of the function $C_1 \cos(\alpha)$ will be $\pm C_1$. Combining the solutions to these first two problems for a data series (call it y) yields

$$y_t = \bar{y} + C_1 \cos\left(\frac{2\pi t}{n}\right). \tag{8.59}$$

This function is plotted as the lighter curve in Fig. 8.12. In this figure the horizon-

Fig. 8.12 Transformation of a simple cosine function defined on 0 to 2π radians to a function representing a data series on the interval 0 to n time units. After changing from time to angular units, multiplying the cosine function by the amplitude C_1 stretches it so that it fluctuates through a range of $2C_1$. Adding the mean of the time series then shifts it to the proper vertical level, producing the lighter curve. The function can then be shifted laterally by subtracting the phase angle ϕ_1 that corresponds to the time of the maximum in the data series (heavier curve).

tal axis indicates the equivalence of angular and time measure, through Eq. (8.57), and the vertical shifting and stretching has produced a function fluctuating around the mean, with a range of $\pm C_1$.

Finally, it is usually necessary to shift a harmonic function laterally in order to have it match the peaks and troughs of a data series. This is most conveniently accomplished when the cosine function is used, because this function achieves its maximum value when the angle on which it operates is zero. Shifting the cosine function to the right by the angle ϕ_1 results in a new function that is maximized at $\omega_1 t/n = \phi_1$,

$$y_t = \bar{y} + C_1 \cos\left(\frac{2\pi t}{n} - \phi_1\right). \tag{8.60}$$

The angle ϕ_1 is called the *phase angle*, or *phase shift*. Shifting the cosine function to the right by this amount requires *subtracting* ϕ_1, so that the argument of the cosine function is zero when $(2\pi t/n) = \phi_1$. Notice that by using Eq. (8.56) it would be possible to rewrite Eq. (8.60) using the sine function. However, the cosine is usually used as in Eq. (8.60), because the phase angle can then be easily interpreted as corresponding to the time of the maximum of the harmonic function. That is, the function in Eq. (8.60) is maximized at time $t = \phi_1 n/2\pi$.

Example 8.8. Transforming a Cosine Wave to Represent the Annual Cycle

Figure 8.13 illustrates the foregoing procedure using the 12 mean monthly temperatures (°F) for 1943–1989 at Ithaca, New York. Figure 8.13a is simply a plot

Fig. 8.13 Approximate matching of a cosine function to a data series. (a) Average monthly temperatures (°F) for Ithaca, New York for the years 1943–1989 (the data values are given in Table 8.5). The annual cycle of average temperature is evidently approximately sinusoidal. (b) Three cosine functions illustrating transformation from time to angular measure (light line at bottom), vertical positioning and stretching (dashed line), and lateral shifting (heavy line) yielding finally the function matching the data approximately. The horizontal dashed lines indicate the average of the 12 data points, 46.1°F.

of the 12 data points, with $t = 1$ indicating January, $t = 2$ indicating February, and so on. The overall annual average temperature of 46.1°F is indicated by the dashed horizontal line. These data appear to be at least approximately sinusoidal, executing a single full cycle over the course of the 12 months. The warmest mean temperature is 68.8°F in July, and the coldest is 22.2°F in January.

The light curve at the bottom of Fig. 8.13b is simply a cosine function with the argument transformed so that it executes one full cycle in 12 months. It is obviously a poor representation of the data. The dashed curve in Fig. 8.13b shows this function lifted to the level of the average annual temperature, and stretched so that its range is similar to that of the data series [Eq. (8.58)]. The stretching has been done only approximately, by choosing the amplitude C_1 to be half the difference between the July and January temperatures.

Finally, the cosine curve needs to be shifted to the right to line up well with the data. The maximum in the curve can be made to occur at $t = 7$ months (July) by introducing the phase shift, using Eq. (8.57), of $\phi_1 = (7)(2\pi)/12 = 7\pi/6$. The result is the heavy curve in Fig. 8.13, which is of the form of Eq. (8.60). This function lines up with the data points, albeit somewhat roughly. The correspondence between the curve and the data can be improved by using better estimators for the amplitude and phase of the cosine wave. ¤

8.4.3 Estimation of the Amplitude and Phase of a Single Harmonic

Although the heavy curve in Fig. 8.13 represents the associated temperature data reasonably well, the correspondence will be improved if better choices for C_1 and ϕ_1 can be found. The easiest way to do this is to use the trigonometric identity

$$\cos(\alpha - \phi_1) = \cos(\phi_1)\cos(\alpha) + \sin(\phi_1)\sin(\alpha). \tag{8.61}$$

Substituting $\alpha = 2\pi t/n$ from Eq. (8.57) and multiplying both sides by the amplitude C_1, we obtain

$$\begin{aligned} C_1 \cos\left(\frac{2\pi t}{n} - \phi_1\right) &= C_1 \cos(\phi_1)\cos\left(\frac{2\pi t}{n}\right) + C_1 \sin(\phi_1)\sin\left(\frac{2\pi t}{n}\right) \\ &= A_1 \cos\left(\frac{2\pi t}{n}\right) + B_1 \sin\left(\frac{2\pi t}{n}\right), \end{aligned} \tag{8.62}$$

where

$$A_1 = C_1 \cos(\phi_1) \tag{8.63a}$$

and

$$B_1 = C_1 \sin(\phi_1). \tag{8.63b}$$

Equation (8.62) says that it is mathematically equivalent to represent a harmonic wave either as a cosine function with amplitude C_1 and phase ϕ_1 or as the sum of an unshifted cosine and unshifted sine wave with amplitudes A_1 and B_1.

For the purpose of estimating one or the other of these pairs of parameters from a set of data, the advantage of representing the wave using the second line of Eq. (8.62) rather than Eq. (8.60) derives from the fact that the former is a linear function of the parameters. Notice that making the variable transformations $x_1 = \cos(2\pi t/n)$ and $x_2 = \sin(2\pi t/n)$, and substituting these into the second line of Eq. (8.62), produces what looks like a two-predictor regression equation with $A_1 = b_1$ and $B_1 = b_2$. In fact, given a data series y_t, one can use this transformation together with ordinary regression software to find least-squares estimates of the parameters A_1 and B_1, with y_t as the predictand. Furthermore, the regression package will also produce the average of the predictand values as the intercept, b_0. Subsequently, the operationally more convenient form of Eq. (8.60) can be recovered by inverting Eqs. (8.63) to yield

$$C_1 = [A_1^2 + B_1^2]^{1/2} \tag{8.64a}$$

and

$$\phi_1 = \begin{cases} \tan^{-1}\dfrac{B_1}{A_1} & & A_1 > 0 \\ \tan^{-1}\dfrac{B_1}{A_1} \pm \pi, & \text{or } \pm 180°, & A_1 < 0 \\ \dfrac{\pi}{2}, \text{ or } 90°, & & A_1 = 0. \end{cases} \tag{8.64b}$$

Notice that since the trigonometric functions are periodic, effectively the same phase angle is produced by adding or subtracting a half-circle of angular measure if $A_1 < 0$. The alternative that produces $0 < \phi_1 < 2\pi$ is usually selected.

Finding the parameters A_1 and B_1 in Eq. (8.62) using least-squares regression will work in the general case. For the special (although not too unusual) situation where the data values are equally spaced in time with no missing values, the properties of the sine and cosine functions allow the same least-squares parameter values to be obtained more easily using

$$A_1 = \frac{2}{n} \sum_{t=1}^{n} y_t \cos\left(\frac{2\pi t}{n}\right) \tag{8.65a}$$

and

$$B_1 = \frac{2}{n} \sum_{t=1}^{n} y_t \sin\left(\frac{2\pi t}{n}\right). \tag{8.65b}$$

Example 8.9. Harmonic Analysis of Average Monthly Temperatures

Table 8.5 shows the calculations necessary to obtain least-squares estimates for the parameters of the annual harmonic representing the Ithaca mean monthly temperatures plotted in Fig. 8.13a, using Eqs. (8.65). The temperature data are shown in the column labeled y_t, and their average is easily computed as

Table 8.5

Mechanics of Using Eq. (8.65) to Estimate Parameters of a Fundamental Harmonic[a]

t	y_t	cos $(2\pi t/12)$	sin $(2\pi t/12)$	$y_t$cos $(2\pi t/12)$	$y_t$sin $(2\pi t/12)$
1	22.2	0.866	0.500	19.225	11.100
2	22.7	0.500	0.866	11.350	19.658
3	32.2	0.000	1.000	0.000	32.200
4	44.4	−0.500	0.866	− 22.200	38.450
5	54.8	−0.866	0.500	− 47.457	27.400
6	64.3	−1.000	0.000	− 64.300	0.000
7	68.8	−0.866	−0.500	− 59.581	−34.400
8	67.1	−0.500	−0.866	− 33.550	−58.109
9	60.2	0.000	−1.000	0.000	−60.200
10	49.5	0.500	−0.866	24.750	−42.867
11	39.3	0.866	−0.500	34.034	−19.650
12	27.4	1.000	0.000	27.400	0.000
Totals:	552.9	0.000	0.000	−110.329	−86.417

[a] The data series y_t are the mean monthly temperatures at Ithaca for month t plotted in Fig. 8.13a. Each of the 12 terms in Eqs. (8.65a) and (8.65b), respectively, are shown in the last two columns.

552.9/12 = 46.1°F. The $n = 12$ terms of the sums in Eqs. (8.65a) and (8.65b) are shown in the last two columns. Application of Eqs. (8.65) to these yields $A_1 = (2/12)(-110.329) = -18.39$, and $B_1 = (2/12)(-86.417) = -14.40$.

Equations (8.64) transform these two amplitudes to the parameters of the amplitude-phase form of Eq. (8.60). This transformation allows easier comparison the heavy curve plotted in Fig. 8.13b. The amplitude is $C_1 = [-18.39^2 + -14.40^2]^{1/2} = 23.36°$F, and the phase angle is $\phi_1 = \tan^{-1}(-14.40/-18.39) + 180° = 218°$. Here 180° has been added rather than subtracted, so that $0° < \phi_1 < 360°$. The least-squares amplitude of $C_1 = 23.36°$F is quite close to the one used to draw Fig. 8.13b, whereas the phase angle is 8° greater than the (7)(360°)/12 = 210° angle that was "eyeballed" on the basis of the July mean being the warmest of the 12 months. The value of $\phi_1 = 218°$ is a better estimate, and implies a somewhat later date for the time of the climatologically warmest temperature at this location. In fact, since there are very nearly as many degrees in a full cycle as there are days in one year, the results from Table 8.5 indicate that the heavy curve in Fig. 8.13b should be shifted to the right by about one week. It is apparent that this would improve the correspondence between the curve and the data points. ¤

Example 8.10. Interpolation of the Annual Cycle to Average Daily Values

The calculations in Example 8.9 result in a smoothly varying representation of the annual cycle of mean temperature at Ithaca, based on the monthly values. Particularly if this were a location for which daily data were not available, it might be valuable to be able to use a function like this to represent the climatological average temperatures on a day-by-day basis. In order to employ the cosine curve in Eq. (8.60) with time t in days, it would be necessary to use $n = 365$ days rather than $n = 12$ months. The amplitude can be left unchanged, although Epstein (1991) suggests a method to adjust this parameter that will produce a somewhat better representation of the annual cycle of daily values. In any case, however, it is necessary to make an adjustment to the phase angle.

Consider that the time $t = 1$ month represents all of January, and thus might be reasonably assigned to the middle of the month, perhaps the 15th. Thus, the $t = 0$ months point of this function corresponds to the middle of December. Therefore, simply substituting the Julian date (January 1 = 1, January 2 = 2, . . . , February 1 = 32, etc.) for the time variable will result in a curve that is shifted too far left by about two weeks. What is required is a new phase angle, say ϕ_1', consistent with a time variable t' in days, that will position the cosine function correctly.

On December 15, the two time variables are $t = 0$ months, and $t' = -15$ days. On December 31, they are $t = 0.5$ month = 15 days, and $t' = 0$ days. Thus, in consistent units, $t' = t - 15$ days, or $t = t' + 15$ days. Substituting $n = 365$ days and $t = t' + 15$ into Eq. (8.60) yields

$$
\begin{aligned}
y_t &= \bar{y} + C_1 \cos\left[\frac{2\pi t}{n} - \phi_1\right] = \bar{y} + C_1 \cos\left[\frac{2\pi(t' + 15)}{365} - \phi_1\right] \\
&= \bar{y} + C_1 \cos\left[\frac{2\pi t'}{365} + 2\pi\frac{15}{365} - \phi_1\right] \\
&= \bar{y} + C_1 \cos\left[\frac{2\pi t'}{365} - \left(\phi_1 - 2\pi\frac{15}{365}\right)\right] \\
&= \bar{y} + C_1 \cos\left[\frac{2\pi t'}{365} - \phi_1'\right].
\end{aligned} \tag{8.66}
$$

That is, the required new phase angle is $\phi_1' = \phi_1 - (15)(2\pi)/365$. ¤

8.4.4 Higher Harmonics

The computations in Example 8.9 produced a single cosine function passing quite close to the 12 monthly mean temperature values. This very good fit results because the shape of the annual cycle of temperature at this location is approximately sinusoidal, with a single full cycle being executed over the $n = 12$ points of the time series. In general, one does not expect that a single harmonic wave will represent every time series this well. However, just as adding more predictors to a multiple regression will improve the fit to a set of dependent data, adding more cosine waves to a harmonic analysis will improve the fit to any time series.

A given data series consisting of n points can be represented exactly, meaning that a harmonic function can be found that passes through each of the points, by adding together a series of $n/2$ harmonic functions,

$$
\begin{aligned}
y_t &= \bar{y} + \sum_{k=1}^{n/2} \left\{ C_k \cos\left[\frac{2\pi kt}{n} - \phi_k\right]\right\} \\
&= \bar{y} + \sum_{k=1}^{n/2} \left\{ A_k \cos\left[\frac{2\pi kt}{n}\right] + B_k \sin\left[\frac{2\pi kt}{n}\right]\right\}.
\end{aligned} \tag{8.67}
$$

Notice that the second line of Eq. (8.67) emphasizes that Eq. (8.62) holds for any cosine wave, regardless of its frequency. The cosine wave constituting the $k = 1$ term on the first line in Eq. (8.67) is simply the fundamental, or first harmonic, that was the subject of the previous section. The other $n/2 - 1$ terms in the summation of Eq. (8.67) are *higher harmonics*, or cosine waves with frequencies

$$
\omega_k = \frac{2\pi k}{n} \tag{8.68}
$$

that are integer multiples of the fundamental frequency ω_1.

For example, the second harmonic is that cosine function that completes exactly two full cycles over the n points of the data series. It has its own amplitude C_2 and

phase angle ϕ_2. Notice that the factor k inside the cosine function in Eq. (8.67) is of critical importance. When $k = 1$, the angle $\alpha = 2\pi kt/n$ varies through a single full cycle of 0 to 2π radians as the time index increased from $t = 0$ to $t = n$, as described above. In the case of the second harmonic where $k = 2$, $\alpha = 2\pi kt/n$ executes one full cycle as t increase to $n/2$, and then executes a second full cycle between $t = n/2$ and $t = n$. Similarly, the third harmonic is defined by the amplitude C_3 and the phase angle ϕ_3, and varies through three cycles as t increases from 0 to n.

The second line of Eq. (8.67) suggests that the coefficients A_k and B_k corresponding to particular data series y_t can be found using multiple regression methods, after the data transformations $x_1 = \cos(2\pi t/n)$, $x_2 = \sin(2\pi t/n)$, $x_3 = \cos(2\pi 2t/n)$, $x_4 = \sin(2\pi 2t/n)$, $x_5 = \cos(2\pi 3t/n)$, and so on. This is, in fact, the case in general, but if the data series is equally spaced in time and contains no missing values, Eqs. (8.65) generalize to

$$A_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(\frac{2\pi kt}{n}\right) \tag{8.69a}$$

and

$$B_k = \frac{2}{n}\sum_{t=1}^{n} y_t \sin\left(\frac{2\pi kt}{n}\right). \tag{8.69b}$$

To compute a particular A_k, for example, these equations indicate than an n-term sum is formed, consisting of the products of the data series y_t with values of a cosine function executing k full cycles during the n time units. For relatively short data series these equations can be easily programmed and evaluated using a hand calculator or spreadsheet software. For larger data series the A_k and B_k coefficients are usually computed using a more efficient method that will be mentioned in Section 8.5. Having computed these coefficients, one can obtain the amplitude-phase form of the first line of Eq. (8.67) by computing, separately for each harmonic,

$$C_k = [A_k^2 + B_k^2]^{1/2} \tag{8.70a}$$

and

$$\phi_k = \begin{cases} \tan^{-1}\dfrac{B_k}{A_k}, & A_k > 0 \\ \tan^{-1}\dfrac{B_k}{A_k} \pm \pi, \quad \text{or} \quad \pm 180°, & A_k < 0 \\ \dfrac{\pi}{2}, \quad \text{or} \quad 90°, & A_k = 0. \end{cases} \tag{8.70b}$$

Recall that a multiple regression function will pass through all the developmental data points and will exhibit $R^2 = 100\%$, if there are as many predictor values as data points. The series of cosine terms in Eq. (8.67) is an instance of this

"overfitting" principle, because there are two parameters (the amplitude and phase) for each harmonic term. Thus the $n/2$ harmonics in Eq. (8.67) consist of n predictor variables, and any set of data, regardless of how "untrigonometric" it may look, can be represented exactly using Eq. (8.67).

Since the sample mean in Eq. (8.67) is effectively one of the estimated parameters, corresponding to the regression intercept b_0, an adjustment to Eq. (8.67) is required if n is odd. In this case a summation over only $(n - 1)/2$ harmonics is required to completely represent the function. That is, $(n - 1)/2$ amplitudes plus $(n - 1)/2$ phase angles plus the sample average of the data equals n. If n is even, there are $n/2$ terms in the summation, but the phase angle for the final and highest harmonic, $\phi_{n/2}$, is zero.

One may or may not want to use all $n/2$ harmonics indicated in Eq. (8.67), depending on the context. Often for defining, say, an annual cycle of a climatological quantity, the first few harmonics may give a quite adequate representation from a practical standpoint. If the goal is to find a function passing exactly through each of the data points, then all $n/2$ harmonics would be used. Recall that Chapter 6 warned against overfitting in the context of developing forecast equations, because the artificial skill exhibited on the developmental data does not carry forward when the equation is used to forecast future independent data. In this latter case the goal would not be to forecast but rather to represent the data, so that the "overfitting" ensures that Eq. (8.67) reproduces a particular data series exactly.

Example 8.11. A More Complicated Annual Cycle

Figure 8.14 illustrates the use of a small number of harmonics to smoothly represent the annual cycle of a climatological quantity. Here the quantity is the

Fig. 8.14 The annual cycle of the climatological probability that no measurable precipitation will fall during the 5-day period centered on the date on the horizontal axis, for El Paso, Texas. Irregular line is the plot of the daily relative frequencies, and the smooth curve is a three-harmonic fit to the data. [From Epstein and Barnston (1988).]

probability (expressed as a percentage) of five consecutive days without measurable precipitation, for El Paso, Texas. The irregular curve is a plot of the individual daily relative frequencies computed using data for the years 1948–1983. These apparently go through a regular annual cycle, with the wettest time of year being summer, and with dry springs and falls separated by a somewhat less dry winter. The figure also shows irregular, short-term fluctuations that have probably arisen mainly from sampling variations particular to the specific years analyzed. If a different sample of El Paso precipitation data had been used to compute the relative frequencies (say, 1900–1935), the same broad pattern would be evident, but the details of the individual "wiggles" would be different.

The annual cycle in Fig. 8.14 is quite evident, yet it does not resemble a simple cosine wave. However, this cycle is reasonably well represented by the smooth curve, which is a three-harmonic representation. That is, the smooth curve is a plot of Eq. (8.67) with $k = 3$. The mean value for this data is 61.4%, and the parameters for the first two of these harmonics are $C_1 = 13.6\%$, $\phi_1 = 72° = 0.4\pi$, $C_2 = 13.8\%$, and $\phi_2 = 272° = 1.51\pi$. These values can be computed from the underlying data using Eqs. (8.69) and (8.70). Computing and plotting the sum of all possible $(365 - 1)/2 = 182$ harmonics would result in a function identical to the irregular curve in Fig. 8.14.

Figure 8.15 illustrates the construction of the smooth curve representing the annual cycle in Fig. 8.14. Panel a shows the first (dashed) and second (solid) harmonics plotted separately, both as a function of time (t) in days and as a function of the corresponding angular measure in radians. Also indicated are the magnitudes of the amplitudes C_k in the vertical, and the correspondence of the phase angles ϕ_k to the maxima of the two functions. Note that since the second harmonic executes two cycles during the full 365 days of the year, there are two times of maximum, located at $\phi_2/2$ and $\pi + \phi_2/2$. (The maxima for the third harmonic would occur at $\phi_3/3$, $2\pi/3 + \phi_3/3$, and $4\pi/3 + \phi_3/3$, with a similar pattern holding for the higher harmonics.)

The curve in Fig. 8.15b is constructed simply by adding the values for the two functions in Fig. 8.15a at each time point. Note that the two panels in Fig. 8.15 have been plotted using different vertical scales. During times of the year where the two harmonics are of opposite sign but comparable magnitude, their sum is near zero. The maximum and minimum of the function in Fig. 8.15b are achieved when its two components have relatively large magnitudes of the same sign. Adding the annual mean value of 61.4% to the lower curve results in a close approximation to the smooth curve in Fig. 8.14, with the small differences between the two attributable to the third harmonic. ¤

8.4.5 Harmonic Dials

The information contained in harmonic functions can be conveniently summarized using a graphical device known as the *harmonic dial.* Harmonic dials are especially useful for displaying the geographic distribution of periodic quantities.

Fig. 8.15 Construction of the smooth curve in Fig. 8.14. (a) The first (dashed) and second (solid) harmonics of the annual cycle are plotted separately. These are defined by $C_1 = 13.6\%$, $\phi_1 = 72° = 0.4\pi$, $C_2 = 13.8\%$, and $\phi_2 = 272° = 1.51\pi$. The horizontal axis is labeled in both days and radians. (b) The smoothed representation of the annual cycle is produced by adding the values of the two functions in panel a for each time point. Subsequently adding the annual mean value of 61.4% produces a curve very similar to that in Fig. 8.14. The small differences are accounted for by the third harmonic. Note that the two panels have different vertical scales.

The harmonic dial depicts the two parameters of a single harmonic using a line segment whose length is proportional to the amplitude C_k, and whose angular orientation (the direction in which it is pointed) indicates the phase angle ϕ_k. Generally a map of harmonic dials will include a legend or key, allowing at least approximate reconstruction of the numerical values of the amplitude and phase for each location. In addition, such a map can convey an appreciation of large-scale spatial trends of the quantity portrayed.

Figure 8.16 shows harmonic dials for the first (a) and second (b) harmonics of the annual cycle of the probability of five consecutive dry days over the United States. Lengths of the symbols are proportional to the amplitudes, and their angular orientations indicate the phases, according to the inset "clock." The arrowheads in panel a point to the date on which the first harmonic is maximized. For example, the arrow for El Paso (32° N, 102° W) points to about "8:00," and is about one-third the length of the inset arrow labeled "35." Figure 8.15a indicates that the amplitude of the first harmonic of this quantity at El Paso is 13.6, which is slightly more than 35/3. Similarly, the phase angle of the first harmonic, locating the maximum of the function, is 72°, or 72/360 of the way through the year. This corresponds approximately to mid-March, as indicated by the inset. There are no arrowheads in Fig. 3.16b since the second harmonic exhibits two maxima, separated by half of a year. The phase angle ϕ_2 for this data at El Paso is 272°, yielding

Fig. 8.16 Harmonic dials representing amplitudes and phases of the first (a) and second (b) harmonics of the annual cycle of the probability of five consecutive dry days. Lengths of the symbols are proportional to the amplitudes, and their angular orientations indicate the phases, according to the inset "clock." The arrowheads in panel a point to date on which the first harmonic is maximized. There are no arrowheads in panel b since the second harmonic exhibits two maxima, separated by half of a year. The symbols at El Paso (32° N, 102° W) summarize the two curves in Fig. 8.15a. [From Epstein and Barnston (1988).]

maxima (cf. Fig. 8.15a) 272/[(2)(360)] = 0.38, or a bit less than four-tenths of the way through the year, corresponding to mid-May. The second hump of the second harmonic occurs 6 months later, in November. The northwest–southeast orientation of the line at El Paso in Fig. 8.16b specifies this time placement of the second harmonic.

In addition to information about the annual cycle of 5-day dry-period probabilities at individual locations, Figure 8.16 allows an appreciation of the differences in the annual cycles of the dry-period climatologies across the United States. As pointed out by Epstein and Barnston (1988), five regimes (or dry-period "climates") can be discerned in the first harmonics portrayed in Fig. 8.16a. In the west and northwest, the long arrows indicate a very pronounced difference between wet and dry seasons, and the orientation of the arrows toward July and August indicate that it is the summers that are dry and the winters that are wet. Locations in the high plains show a late fall or early winter maximum in dry-period probabilities, with the difference between dry and wet times of the year (lengths of arrows) less pronounced than on the west coast. Most of the eastern United States shows a slight tendency for higher probabilities of dry periods in the fall; except Florida, where the first harmonic is greatest in late winter. Stations in the desert southwest exhibit low-amplitude (weak) cycles, fluctuating around a rather large mean probability. Locations in the southeast and southwest for which the amplitude of the second harmonic is comparable to or larger than that for the first harmonic apparently exhibit a tendency for two relatively dry seasons and two relatively wet seasons per year.

Notice that the amplitudes and phases for El Paso in the two panels of Fig. 8.16 do not fully define the annual cycles at these locations. The missing information is the annual average, which allows correct vertical placement of the smooth function defined by the sum of the first two harmonics. For example, the amplitude at El Paso in Fig. 8.16a indicates that the first harmonic varies between the annual mean plus 13.6% and the annual mean minus 13.6%, but does not locate the mean. A more complete display of the geographic distribution this climatological quantity would include a third map, indicating the spatial variation of the annual mean probability.

8.4.6 The Harmonic Functions as Uncorrelated Regression Predictors

The second line of Eq. (8.67) suggests the use of multiple regression to find best-fitting harmonics for a given data series y_t. But for equally spaced data with no missing values, Eqs. (8.69) will produce the same least-squares estimates for the coefficients A_k and B_k as will multiple regression software. Notice, however, that Eqs. (8.69) do not depend on any harmonic other than the one whose coefficients are being computed. That is, these equations depend on the current value of k, but not $k - 1$, or $k - 2$, or any other harmonic index. This fact implies that the coefficients A_k and B_k for any particular harmonic can be computed independently of those for any other harmonic.

Recall that usually regression parameters need to be recomputed each time a new predictor variable is entered into a multiple regression equation, and each time a predictor variable is removed from a regression equation. As noted in Chapter 6, this recomputation is necessary in the general case of sets of predictor variables that are mutually correlated, because correlated predictors carry redundant information to a greater or lesser extent. It is a remarkable property of the harmonic functions that they are uncorrelated so that, for example, the parameters (amplitude and phase) for the first or second harmonic are the same regardless of whether they will be used in an equation with the third, fourth, or any other harmonics.

This remarkable attribute of the harmonic functions is a consequence of what is called the *orthogonality property* of the sine and cosine functions. That is, for integer harmonic indices k and j,

$$\sum_{t=1}^{n} \cos\left(\frac{2\pi kt}{n}\right)\sin\left(\frac{2\pi jt}{n}\right) = 0, \quad \text{for any integer values of } k \text{ and } j; \tag{8.71a}$$

and

$$\sum_{t=1}^{n} \cos\left(\frac{2\pi kt}{n}\right)\cos\left(\frac{2\pi jt}{n}\right) = \sum_{t=1}^{n} \sin\left(\frac{2\pi kt}{n}\right)\sin\left(\frac{2\pi jt}{n}\right) = 0, \quad \text{for} \quad k \neq j. \tag{8.71b}$$

Consider, for example, the two transformed predictor variables $x_1 = \cos[2\pi t/n]$ and $x_3 = \cos[2\pi(2t)/n]$. The Pearson correlation between these derived variables is given by

$$r_{x_1x_3} = \frac{\Sigma_{t=1}^{n} (x_1 - \bar{x}_1)(x_3 - \bar{x}_3)}{[\Sigma_{t=1}^{n} (x_1 - \bar{x}_1)^2 \, \Sigma_{t=1}^{n} (x_3 - \bar{x}_3)^2]^{1/2}}, \tag{8.72a}$$

and since the averages $\bar{x}_1$ and $\bar{x}_3$ of cosine functions over integer numbers of cycles are zero,

$$r_{x_1x_3} = \frac{\Sigma_{t=1}^{n} \cos(2\pi t/n)\cos(2\pi 2t/n)}{[\Sigma_{t=1}^{n} \cos^2(2\pi t/n) \, \Sigma_{t=1}^{n} \cos^2(2\pi 2t/n)]^{1/2}} = 0, \tag{8.72b}$$

because the numerator is zero by Eq. (8.71b).

Since the relationships between harmonic predictor variables and the data series y_t do not depend on what other harmonic functions are also being used to represent the series, the proportion of the variance of y_t accounted for by each harmonic is also fixed. When this proportion is expressed as the R^2 statistic commonly computed in regression, the R^2 for the kth harmonic is simply

$$R_k^2 = \frac{n/2 \; C_k^2}{(n-1)s_y^2}. \tag{8.73}$$

In terms of the regression ANOVA table, the numerator of Eq. (8.73) is the regression sum of squares for the kth harmonic. The factor s_y^2 is simply the sample

variance of the data series, so the denominator of Eq. (8.73) is the total sum of squares, SST. Notice that the strength of the relationship between the kth harmonic and the data series is expressed entirely by the amplitude C_k. The phase angle ϕ_k is necessary only to determine the positioning of the cosine curve in time. Furthermore, since each harmonic provides independent information about the data series, the joint R^2 exhibited by a regression equation with only harmonic predictors is simply the sum of the R_k^2 values for each of the harmonics,

$$R^2 = \sum_{k \text{ in the equation}} R_k^2. \tag{8.74}$$

If all the $n/2$ possible harmonics are used as predictors [Eq. (8.67)], then the total R^2 in Eq. (8.74) will be exactly 1.

8.5 Frequency Domain. II. Spectral Analysis

Equation (8.67) says that a data series y_t of length n can be completely specified in terms of the n parameters of $n/2$ harmonic functions. Equivalently, one can take the view that the data series y_t is transformed into new set of quantities A_k and B_k according to Eqs. (8.69). For this reason, Eqs. (8.69) are called the *discrete Fourier transform.* Equivalently, the data series can be represented as the n quantities C_k and ϕ_k, obtained from the A_k and B_k values using the transformations in Eqs. (8.70). According to Eq. (8.74), this data transformation accounts for all the variation in the series y_t.

The foregoing suggests that a different way to look at a time series is as a collection of Fourier coefficients A_k and B_k that are a function of frequency ω_k [Eq. (8.68)], rather than as a collection of data points y_t measured as a function of time. The advantage of this new perspective is that it allows one to see separately the contributions to a time series that are made by processes varying at different speeds; that is, by processes operating at a spectrum of different frequencies. Panofsky and Brier (1958) illustrate this distinction with a nice analogy: "An optical spectrum shows the contributions of different wave lengths or frequencies to the energy of a given light source. The spectrum of a time series shows the contributions of oscillations with various frequencies to the variance of a time series." Even if the underlying physical basis for a data series y_t is not really well represented by a series of cosine waves, often much can still be learned about the data by viewing it from this perspective.

8.5.1 The Periodogram or Fourier Line Spectrum

The characteristics of a time series that has been Fourier-transformed into the frequency domain are most often examined graphically, using a plot known as the *periodogram,* or *Fourier line spectrum.* This plot is sometimes also called the *power spectrum,* or simply *the spectrum* of the data series.

In simplest form, this plot of a spectrum consists of the squared amplitudes C_k^2 as a function of the frequencies ω_k. The vertical axis is sometimes numerically rescaled, in which case the plotted points are proportional to the squared amplitudes. One choice for this proportional rescaling is that in Eq. (8.73). Note that information contained in the phase angles ϕ_k is not portrayed in the spectrum. Therefore, the spectrum conveys the proportion of variation in the original data series accounted for by oscillations at the harmonic frequencies, but does not supply information about when in time these oscillations are expressed.

It is common for the vertical axis of a spectrum to be plotted on a logarithmic scale. Plotting the vertical axis logarithmically is particularly useful if the variations in the time series are dominated by harmonics of only a few frequencies. In this case a linear plot would result in the remaining spectral components being invisibly small. A logarithmic vertical axis also regularizes the representation of confidence limits for the spectral estimates, as will be illustrated later.

The horizontal axis of the line spectrum consists of $n/2$ frequencies ω_k if n is even, and $(n - 1)/2$ frequencies if n is odd. The smallest of these will be the lowest frequency $\omega_1 = 2\pi/n$ (the fundamental frequency), and this corresponds to the cosine wave that executes a single cycle over the n time points. The highest frequency, $\omega_{n/2} = \pi$, is called the *Nyquist frequency*. It is the frequency of the cosine wave that executes a full cycle over only two time intervals, and that executes $n/2$ cycles over the full data record. The Nyquist frequency depends on the time resolution of the original data series y_t, and imposes an important limitation on the information available from a spectral analysis.

The horizontal axis is often simply the angular frequency, ω, with units of radians/time. A common alternative is to use the frequencies

$$f_k = \frac{k}{n} = \frac{\omega_k}{2\pi}, \tag{8.75}$$

which have dimensions of time^{-1}. Under this alternative convention, the allowable frequencies range from $f_1 = 1/n$ for the fundamental to $f_{n/2} = 1/2$ for the Nyquist frequency. The horizontal axis of a spectrum can also be scaled according to the reciprocal of the frequency, or the period of the kth harmonic

$$\tau_k = \frac{n}{k} = \frac{2\pi}{\omega_k} = \frac{1}{f_k}. \tag{8.76}$$

The period τ_k specifies the length of time required for a cycle of frequency ω_k to be completed. Associating periods with the periodogram estimates can help one visualize the time scales on which the important variations in the data are occurring.

Example 8.12. Discrete Fourier Transform of a Small Data Set

Table 8.6 shows a simple data set and its discrete Fourier transform. Columns 1–3 list the observed average monthly temperatures at Ithaca, New York, for the

Table 8.6

Average Monthly Temperatures (°F) at Ithaca, New York, for 1987–1988[a]

Month	1987	1988	k	τ_k, (months)	A_k	B_k	C_k
1	21.4	20.6	1	24.00	− 0.14	0.44	0.46
2	17.9	22.5	2	12.00	−23.76	−2.20	23.86
3	35.9	32.9	3	8.00	−0.99	0.39	1.06
4	47.7	43.6	4	6.00	−0.46	−1.25	1.33
5	56.4	56.5	5	4.80	−0.02	−0.43	0.43
6	66.3	61.9	6	4.00	−1.49	−2.15	2.62
7	70.9	71.6	7	3.43	−0.53	−0.07	0.53
8	65.8	69.9	8	3.00	−0.34	−0.21	0.40
9	60.1	57.9	9	2.67	1.56	0.07	1.56
10	45.4	45.2	10	2.40	0.13	0.22	0.26
11	39.5	40.5	11	2.18	0.52	0.11	0.53
12	31.3	26.7	12	2.00	0.79	—	0.79

[a] The sample mean of these data is 46.1°F, and the sample standard deviation is 17.46°F. Columns 4–8 present the discrete Fourier transform of the $n = 24$ data values.

two years 1987 and 1988. This is such a familiar type of data that, even without doing a spectral analysis, one knows in advance that the primary feature will be the annual cycle of cold winters and warm summers. This expectation is validated by the plot of the data in Fig. 8.17a, which shows these temperatures as a function of time. The overall impression is of a data series that is approximately sinusoidal with a period of 12 months, but that a single cosine wave with this period would not pass exactly through all the points.

Columns 4–8 in Table 8.6 show the same data after being subjected to the discrete Fourier transform. Since $n = 24$ is an even number, the data are completely represented by $n/2 = 12$ harmonics. These are indicated by the rows labeled by the harmonic index, k. Column 5 in Table 8.6 indicates the period [Eq. (8.76)] of each of the 12 harmonics used to represent the data. The period of the fundamental frequency, $\tau_1 = 24$ months, is equal to the length of the data record. Since there are two annual cycles in the $n = 24$ month record, it is the $k =$ 2nd harmonic with period $\tau_2 = 24/2 = 12$ months that is expected to be most important. The Nyquist frequency is $\omega_{12} = \pi$ radians per month (rad/month), or $f_{12} = 0.5$ month^{-1}, corresponding to the period $\tau_{12} = 2$ months.

The coefficients A_k and B_k that would be used in Eq. (8.67) to reconstruct the original data are shown in the next columns of the table. These constitute the discrete Fourier transform of the data series of temperatures. Notice that there are only 23 Fourier coefficients, because 24 independent pieces of information are necessary to fully represent the $n = 24$ data points, including the sample mean of 46.1°F. To use Eq. (8.67) to reconstitute the data, one would substitute $B_{12} = 0$.

Column 8 in Table 8.6 shows the amplitudes C_k, computed according to Eq. (8.70a). The phase angles could also be computed, using Eq. (8.70b), but these

Fig. 8.17 Relationship between a simple time series and its spectrum. (a) Average monthly temperatures at Ithaca, New York for 1987–1988, from Table 8.6. The data are approximately sinusoidal, with a period of 12 months. (b) The spectrum of the data series in panel a, plotted in the form of a histogram, and expressed in the normalized form of Eq. (8.66). Clearly, the most important variations in the data series are represented by the second harmonic, with period $\tau_2 = 12$ months, which is the annual cycle. Note that the vertical scale is logarithmic, so that the next most important harmonic accounts for barely more than 1% of the total variation. The horizontal scale is linear in frequency.

are not needed to plot the spectrum. Figure 8.17b shows the spectrum for these temperature data, plotted in the form of a histogram. The vertical axis consists of the squared amplitudes C_k^2, normalized according to Eq. (8.73) to show the R^2 attributable to each harmonic. The horizontal axis is linear in frequency, but the corresponding periods are also shown, to aid the interpretation. Clearly, most of the variation in the data is described by the second harmonic, the R^2 for which is 97.5%. As expected, the variations of the annual cycle dominate these data, but the fact that the amplitudes of the other harmonics are not zero indicates that the data do not consist of a pure cosine wave with a frequency of $f_2 = 1$ year^{-1}.

Notice, however, that the logarithmic vertical axis tends to deemphasize the smallness of these other harmonics. If the vertical axis were scaled linearly, the plot would consist of a spike at $k = 2$, a small bump at $k = 6$, with the rest of the points being essentially indistinguishable from the horizontal axis. ¤

Example 8.13. Another Sample Spectrum

A less trivial example spectrum is shown in Fig. 8.18. This is a portion of the spectrum of the monthly Tahiti minus Darwin sea-level pressure time series for 1951–1979. That time series resembles the (normalized) SOI index shown in Fig. 3.10, including the tendency for a quasiperiodic behavior. That the variations in the time series are not strictly periodic is evident from the irregular variations in Fig. 3.10, and from the broad (i.e., spread over many frequencies) maximum in the spectrum. Figure 8.18 indicates that the typical length of one of these pseudocycles (corresponding to typical times between El Niño events) is something between $\tau = [(1/36)\ \text{month}^{-1}]^{-1} = 3$ years and $\tau = [(1/84)\ \text{month}^{-1}]^{-1} = 7$ years.

The vertical axis in Fig. 8.18 has been plotted on a linear scale, but units have been omitted because they do not contribute to a qualitative interpretation of the plot. The horizontal axis is linear in frequency, and therefore nonlinear in period. Notice also that the labeling of the horizontal axis indicates that the full spectrum of the underlying data series is not presented in the figure. Since the data series consists of monthly values, the Nyquist frequency must be 0.5 month^{-1}, corresponding to a period of 2 months. Only the left-most one-eighth of the spectrum

Fig. 8.18 The low-frequency portion of the smoothed spectrum for the monthly time series of Tahiti minus Darwin sea-level pressures, 1951–1979. This underlying time series resembles that in Fig. 3.10, and the tendency for oscillations to occur in roughly 3–7-year cycles is reflected in the broad maximum of the spectrum in this range. [After Chen (1982a).]

has been shown because it is these lower frequencies that reflect the physical phenomenon of interest, namely, the El Niño–Southern Oscillation (ENSO) cycle. The estimated spectral density function for the omitted higher frequencies would only create a long, irregular and generally uninformative right tail. ¤

8.5.2 Computing Spectra

One way to compute the spectrum of a data series is to simply apply Eq. (8.69), and then to find the amplitudes using Eq. (8.70b). This is a reasonable approach for relatively short data series and can be programmed easily using, for example, spreadsheet software. These equations would be implemented only for $k = 1, 2, \ldots, (n/2 - 1)$. Because one wants exactly n Fourier coefficients (A_k and B_k) to represent the n points in the data series, the computation for the highest harmonic, $k = n/2$, is done using

$$A_{n/2} = \begin{cases} \left(\frac{1}{2}\right)\left(\frac{2}{n}\right)\sum_{t=1}^{n} y_t \cos\left[\frac{2\pi(n/2)t}{n}\right] = \frac{1}{n}\sum_{t=1}^{n} y_t \cos[\pi t], & n \text{ even} \\ 0, & n \text{ odd} \end{cases} \tag{8.77a}$$

$$B_{n/2} = 0, \qquad n \text{ even or odd.} \tag{8.77b}$$

Although straightforward notationally, this method of computing the discrete Fourier transform is quite inefficient computationally. In particular, many of the calculations called for by Eqs. (8.69) are redundant. Consider, for example, the data for April 1987 in Table 8.6. The term for $t = 4$ and $k = 1$ in the summation in Eq. (8.69b) is $(47.7°\text{F})\sin[(2\pi)(1)(4)/24] = (47.7°\text{F})(0.866) = 41.31°\text{F}$. However, the term involving this same data point for $k = 2$ is exactly the same: $(47.7°\text{F})\sin[(2\pi)(2)(4)/24] = (47.7°\text{F})(0.866) = 41.31°\text{F}$. There are many other such redundancies in the computation of discrete Fourier transforms using Eqs. (8.69). These can be avoided through the use of clever algorithms known as *fast Fourier transforms* (FFTs). Most packages of scientific computational software include one or more FFT routines, which give very substantial speed improvements, especially as the length of the data series increases. In comparison to computation of the Fourier coefficients using a regression approach, an FFT is approximately $n/\log_2(n)$ times faster; or about 15 times faster for $n = 100$, and about 750 times faster for $n = 10{,}000$.

It is worth noting that FFTs are usually documented and implemented in terms of the Euler complex exponential notation,

$$e^{i\omega t} = \cos(\omega t) - i\sin(\omega t), \tag{8.78}$$

where i is the unit imaginary number satisfying $i = \sqrt{-1}$, and $i^2 = -1$. Complex exponentials are used rather than sines and cosines purely as a notational

convenience that makes some of the manipulations less cumbersome. The mathematics are still entirely the same. In terms of complex exponentials, Eq. (8.67) becomes

$$y_t = \bar{y} + \sum_{k=1}^{n/2} H_k \, e^{i[2\pi k/n]t}, \tag{8.79}$$

where H_k is the complex Fourier coefficient

$$H_k = A_k + i \, B_k. \tag{8.80}$$

That is, the real part of H_k is the coefficient A_k, and the imaginary part of H_k is the coefficient B_k.

8.5.3 Aliasing

Aliasing is a hazard inherent in spectral analysis of discrete data. It arises because of the limits imposed by the sampling interval, or the time between pairs of data points. Since a minimum of two points are required to even sketch a cosine wave—one point for the peak and one point for the trough—the highest representable frequency is the Nyquist frequency, with $\omega_{n/2} = \pi$, or $f_{n/2} = 0.5$. A wave of this frequency executes once cycle every two data points, and thus a discrete data set can represent explicitly variations that occur no faster than this speed.

It is worth wondering what happens to the spectrum of a data series if it includes important physical process that vary faster than the Nyquist frequency. If so, the data series is said to be *undersampled*, which means that the points in the time series are spaced too far apart to properly represent these fast variations. However, variations that occur at frequencies higher than the Nyquist frequency do not disappear. Rather, their contributions are spuriously attributed to some lower but representable frequency, between ω_1 and $\omega_{n/2}$. These high-frequency variations are said to be aliased.

Figure 8.19 illustrates the meaning of aliasing. Imagine that the physical data-generating process is represented by the dashed cosine curve. The data series y_t is produced by sampling this process at integer values of the time index t, resulting in the points indicated by the heavy circles. However, the frequency of the dashed curve ($\omega = 8\pi/5$, or $f = 4/5$) is higher than the Nyquist frequency ($\omega = \pi$, or $f = 1/2$), meaning that it oscillates too quickly to be adequately sampled at this time resolution. Rather, if only the information in the discrete time points is available, these data look like the heavy cosine function, the frequency of which ($\omega = 2\pi/5$, or $f = 1/5$) is lower than the Nyquist frequency, and are therefore representable. Note that because the cosine functions are orthogonal, this same effect will occur regardless of whether variations of different frequencies are also present in the data.

The effect of aliasing on spectral analysis is that any "energy" (squared amplitudes) attributable to processes varying at frequencies higher than the Nyquist

Fig. 8.19 Plot depicting the basis of aliasing. Heavy circles represent data points in the data series y_t. Fitting a harmonic function to them results in the heavy curve. However, if the data series had actually been produced by the process indicated by the light dashed curve, the fitted heavy curve would present the misleading impression that the source of the data was actually fluctuating at the lower frequency. The lighter curve has not been sampled densely enough because its frequency, $\omega = 8\pi/5$ (or $f = 4/5$), is higher than the Nyquist frequency of $\omega = \pi$ (or $f = 1/2$). Variations at the frequency of the dashed curve are said to be aliased into the frequency of the heavier curve.

frequency will be erroneously added to that of one of the $n/2$ frequencies that are represented by the spectrum. A frequency $f_A > 1/2$ will be aliased into one of the representable frequencies f (with $0 < f \leq 1/2$) if it differs by an integer multiple of 1 time^{-1}, that is, if

$$f_A = j \pm f, \qquad j = \text{any positive integer.} \tag{8.81a}$$

In terms of angular frequency, variations at the aliased frequency ω_A appear to occur at the representable frequency ω if

$$\omega_A = 2\pi j \pm \omega, \qquad j = \text{any positive integer.} \tag{8.81b}$$

These equations imply that the squared amplitudes for frequencies higher than the Nyquist frequency will be added to those of the representable frequencies in an accordion-like pattern, with each "fold" of the accordion occurring at an integer multiple of the Nyquist frequency. For this reason the Nyquist frequency is sometimes called the *folding frequency*. An aliased frequency f_A that is just slightly higher than the Nyquist frequency of $f_{n/2} = 1/2$ is aliased to a frequency slightly lower than 1/2. Frequencies only slightly lower than twice the Nyquist frequency are aliased to frequencies only slightly higher than zero. The pattern then reverses for $2f_{n/2} < f_A < 3f_{n/2}$. That is, frequencies just higher than $2f_{n/2}$ are aliased to very low frequencies, and frequencies almost as high as $3f_{n/2}$ are aliased to frequencies near $f_{n/2}$.

Figure 8.20 illustrates the effects of aliasing on a hypothetical spectrum. The lighter line represents the true spectrum, which exhibits a concentration of density at low frequencies, but also has a sharp peak at $f = 5/8$ and a broader peak at $f = 19/16$. These second two peaks occur at frequencies higher than the Nyquist fre-

Fig. 8.20 Aliasing in a hypothetical spectrum. The true spectrum (lighter line) exhibits a sharp peak at $f = 5/8$, and a broader peak at $f = 19/16$. Since both of these frequencies are higher than the Nyquist frequency $f = 1/2$, they are aliased in the spectrum (erroneously attributed) to the frequencies indicated. The aliasing follows an accordion-like pattern, with the area between the light line and the dashed line contributed by frequencies from $f = 1$ to $f = 1/2$, and the area between the dashed line and the heavy line contributed by frequencies between $f = 1$ and $f = 3/2$. The resulting apparent spectrum (heavy line) includes both the true spectral density values for frequencies between zero and 1/2, as well as the contributions from the aliased frequencies.

quency of $f = 1/2$, which means that the physical process that generated the data was not sampled often enough to resolve them explicitly. The variations actually occurring at the frequency $f_A = 5/8$ are aliased to (i.e., appear to occur at) the frequency $f = 3/8$. That is, according to Eq. (8.81a), $f_A = 1 - f = 1 - 3/8 = 5/8$. In the spectrum, the squared amplitude for $f_A = 5/8$ is added to the (genuine) squared amplitude at $f = 3/8$ in the true spectrum. Similarly, the variations represented by the broader hump centered at $f_A = 19/16$ in the true spectrum are aliased to frequencies at and around $f = 3/16$ ($f_A = 1 + f = 1 + 3/16 = 19/16$). The dashed line in Fig. 8.20 indicates the portions of the aliased spectral energy (the total area between the light and dark lines) contributed by frequencies between $f = 1/2$ and $f = 1$ (the area below the dashed line), and by frequencies between $f = 1$ and $f = 3/2$ (the area above the dashed line), emphasizing the "fanfolded" nature of the aliased spectral density.

Unfortunately, once a data series has been collected, there is no way to "dealias" its spectrum. That is, it is not possible to tell from the data values alone whether appreciable contributions to the spectrum have been made by frequencies higher than $f_{n/2}$, or how large these contributions might be. In practice, it is desir-

able to have an understanding of the physical basis of the processes generating the data series, so that one can see in advance that the sampling rate is adequate. Of course, in an exploratory setting this advice is of no help, since there the point of the analysis is exactly to gain a better understanding of an unknown or a poorly known generating process. In this latter situation, one would like to see the spectrum approach zero for frequencies near $f_{n/2}$, which would give some hope that the contributions from higher frequencies are minimal. A spectrum such as the heavy line in Fig. 8.20 would lead one to expect that aliasing might be a problem, since its not being essentially zero at the Nyquist frequency could well mean that the true spectrum is nonzero at higher frequencies as well.

8.5.4 Sampling Properties of Spectral Estimates

Since the data from which atmospheric spectra are computed are subject to sampling fluctuations, the Fourier coefficients computed from these data will exhibit random batch-to-batch variations as well. That is, different data batches from the same source will transform to somewhat different C_k^2 values, resulting in somewhat different sample spectra.

Each squared amplitude is an unbiased estimator of the true spectral density, which means that averaged over many batches the mean of the many C_k^2 values would closely approximate the true population value. Unfortunately, the sampling distribution for an individual C_k^2 is rather broad. The sampling distribution of suitably scaled squared amplitudes is the χ^2 distribution with $\nu = 2$ degrees of freedom, which is a gamma distribution having $\alpha = 1$ (cf. Fig. 4.6). A salient property of this distribution is that its standard deviation is equal to its mean. One consequence of this is that the widths of computed confidence intervals are proportional to the individual C_k^2 estimates, so that confidence intervals for spectral estimates plotted with logarithmic vertical axes are the same for all frequencies.

A less happy consequence of the equality of standard deviation and mean for this sampling distribution is that the individual C_k^2 estimates represent the true spectrum rather poorly. The very erratic nature of raw spectral estimates is illustrated by the two sample spectra shown in Fig. 8.21. The solid and lightface lines are two sample spectra computed from different batches of $n = 30$ independent Gaussian random variables. Each of the two sample spectra varies rather wildly around the true spectrum, which is shown by the dashed horizontal line. In a real application, the true spectrum is, of course, not known in advance, and Fig. 8.21 shows that the poor sampling properties of the individual spectral estimates can make it very difficult to discern much about the true spectrum if only a single sample spectrum is available.

The usual remedy in statistics for an unacceptably broad sampling distribution is to increase the sample size. For spectra, however, simply increasing the sample size does not give more precise information about any of the individual frequencies, but rather results in equally imprecise information about more frequencies. For example, the spectra in Fig. 8.21 were computed from $n = 30$ data points, and

Fig. 8.21 The erratic sampling characteristics of estimated spectra. The solid and lightface curves are two sample spectra, each computed using different batches of $n = 30$ independent Gaussian random variables. Both are quite erratic, and points of agreement are more fortuitous than meaningful. The true spectrum for the underlying serially independent data is shown by the horizontal dashed line. The vertical axis is linear.

thus consist of $n/2 = 15$ squared amplitudes. Doubling the sample size to $n = 60$ data values would result in a spectrum at $n/2 = 30$ frequencies, each point of which would exhibit the same large sampling variations as the individual C_k^2 values in Fig. 8.21.

It is possible, however, to use larger data samples to obtain sample spectra that are more representative of the underlying population spectra. One approach is to compute replicate spectra from separate sections of the time series, and then to average the resulting squared amplitudes. In the context of Fig. 8.21, for example, a time series of $n = 60$ could be split into two series of length $n = 30$. The two spectra in Fig. 8.21 might be viewed as having resulted from such a process. Here averaging each of the $n/2 = 15$ pairs of C_k^2 values would result in a less erratic spectrum, that somewhat more faithfully represents the true spectrum. In fact, the sampling distributions of each of these $n/2$ average spectral values would be proportional to the χ^2 distribution with $\nu = 4$, or a gamma distribution with $\alpha = 2$. This distribution is substantially less variable, with standard deviations of $1/\sqrt{2}$ of the averaged estimates, or about 70% of the previous individual estimates. If one had a data series with $n = 300$ points, 10 sample spectra could be computed whose average would smooth out a large fraction of the sampling variability evident in Fig. 8.21. The sampling distribution in this case would have $\nu = 20$, or $\alpha = 10$. The standard deviations for these averages would be smaller by the factor $1/\sqrt{10}$, or about one-third of the magnitudes of the individual averaged squared amplitudes. Since the confidence interval widths are still proportional to the estimated squared amplitudes, a logarithmic vertical scale again results in plotted confidence interval widths not depending on frequency.

An essentially equivalent approach to obtaining a smoother and more representative spectrum using more data begins with computation of the discrete Fourier transform for the longer data series. Initially this results in more highly variable spectral estimates; the sampling variability can be smoothed out by adding (not averaging) the squared amplitudes for groups of adjacent frequencies. The spectrum shown in Fig. 8.18 has been smoothed in this way. For example, if one wanted to estimate the spectrum at the 15 frequencies plotted in Fig. 8.21, these could be obtained by summing pairs of the 30 squared amplitudes obtained from the spectrum of a data record that was $n = 60$ observations long. If $n = 300$ observations were available, the spectrum at these same 15 frequencies could be estimated by adding the squared amplitudes for groups of 10 of the $n/2 = 150$ original frequencies. Here again the sampling distribution is χ^2 with ν equal to twice the number of pooled frequencies, or gamma with α equal to the number of pooled frequencies. Again, logarithmic confidence interval widths would not depend on frequency.

Jenkins and Watts (1968) present a number of smoothing functions that are commonly applied to sample spectra. Note that for any of these, the increased smoothness and representativeness of the resulting spectra come at the expense of decreased frequency resolution. Essentially, stability of the sampling distributions of the spectral estimates is obtained by "smearing" spectral information from a range of frequencies across a frequency band. Smoothing across broader bands produces less erratic estimates, but hides sharp contributions made at particular frequencies. In practice, some compromise between stability and resolution is needed, which is resolved as a matter of subjective judgment.

8.5.5 Theoretical Spectra of Autoregressive Models

Another perspective on the time-domain autoregressive models described in Section 8.3 is provided by their spectra. The types of time dependence produced by different autoregressive models produce characteristic spectral signatures that can be related to the autoregressive parameters.

The simplest case is the AR(1) process, Eq. (8.16). Here positive values of the single autoregressive parameter ϕ induce a memory into the time series that tends to smooth over short-term (high-frequency) variations in the ϵ series, and emphasize the slower (low-frequency) variations. In terms of the spectrum, these effects lead to more density at lower frequencies, and less density at higher frequencies. Furthermore, these effects are progressively stronger for ϕ closer to 1.

These ideas are quantified by the theoretical spectral density function for AR(1) processes:

$$S(f) = \frac{4\sigma_\epsilon^2/n}{1 + \phi^2 - 2\phi \cos(2\pi f)}, \qquad 0 \le f \le 1/2. \tag{8.82}$$

This is a function that associates a spectral density with all frequencies in the range $0 \le f \le 1/2$. The shape of the function is determined by the denominator, whereas

the numerator contains scaling constants that give the function numerical values that are comparable to the empirical spectrum of squared amplitudes, C_k^2. This equation also applies to negative values of the autoregressive parameter, which produce time series tending to oscillate quickly around the mean, and for which the spectral density is greatest at the high frequencies.

Note that, for zero frequency, Eq. (8.82) is proportional to the variance of a time average. This can be appreciated by substituting $f = 0$, substituting Eq. (8.21) and (8.39) into Eq. (8.82), and comparison to Eq. (8.36). Thus, the extrapolation of the spectrum to zero frequency has been used to estimate variances of time averages (e.g., Madden and Shea, 1978).

Figure 8.22 shows theoretical spectra for the AR(1) processes having $\phi = 0.5$,

Fig. 8.22 Theoretical spectral density functions for four AR(1) processes, computed using Eq. 8.82. Autoregressions with $\phi > 0$ are "red noise" processes, since their spectra are enriched at the lower frequencies and depleted at the higher frequencies. The spectrum for the $\phi = 0$ process (i.e., serially independent data) is flat, exhibiting no tendency to emphasize either high- or low-frequency variations. This is a "white noise" process. The autoregression with $\phi = -0.6$ tends to exhibit rapid variations around its mean, which results in a spectrum enriched in the high frequencies, or a "blue noise" process. Autocorrelation functions for the $\phi = 0.5$ and $\phi = -0.6$ processes are shown as insets in Fig. 8.7.

0.3, 0.0, and -0.6. The autocorrelation functions for the first and last of these are shown as insets in Fig. 8.7. As could be anticipated, the two processes with $\phi > 0$ show enrichment of the spectra in the lower frequencies and depletion of the spectra in the higher frequencies, and these characteristics are stronger for the process with the larger autoregressive parameter. By analogy to the properties of visible light, AR(1) processes with $\phi > 0$ are sometimes referred to as "red noise" processes.

The AR(1) process with $\phi = 0$ actually consists of the series of temporally uncorrelated data values $X_{t+1} = \mu + \epsilon_{t+1}$ [cf. Eq. (8.16)]. These exhibit no tendency to emphasize either low-frequency or high-frequency variations, so their spectrum is constant, or flat. Again by analogy to visible light, this is called "white noise" because of the equal mixture of all frequencies. This analogy is the basis of the independent series of ϵ values being called the *white-noise forcing*, and of the parameter σ_ϵ^2 being known as the *white-noise variance*.

Finally, the AR(1) process with $\phi = -0.6$ tends to produce erratic short-term variations in the time series, resulting in negative correlations at odd lags and positive correlations at even lags. (This kind of correlation structure is rare in atmospheric time series.) The spectrum for this process is thus enriched at the high frequencies and depleted at the low frequencies, as indicated in Fig. 8.22. Such series are accordingly known as "blue noise" processes.

Expressions for the spectra of other autoregressive processes, and for ARMA processes as well, are given in Box and Jenkins (1976). Of particular importance is the spectrum for the AR(2) process:

$$S(f) = \frac{4\sigma_\epsilon^2/n}{1 + \phi_1^2 + \phi_2^2 - 2\phi_1(1 - \phi_2)\cos(2\pi f) - 2\phi_2 \cos(4\pi f)},$$

$$0 \le f \le 1/2. \tag{8.83}$$

This equation reduces to Eq. (8.82) for $\phi_2 = 0$, since an AR(2) process [Eq. (8.27)] with $\phi_2 = 0$ is simply an AR(1) process.

The AR(2) processes are particularly interesting because of their capacity to exhibit a wide variety of behaviors, including pseudoperiodicities. This diversity is reflected in the various forms of the spectra that are included in Eq. (8.83). Figure 8.23 illustrates a few of these, corresponding to the AR(2) processes whose autocorrelation functions are shown as insets in Fig. 8.7. The processes with $\phi_1 = 0.9$, $\phi_2 = -0.6$ and $\phi_1 = -0.9$, $\phi_2 = -0.5$ exhibit pseudoperiodicities, as indicated by the broad humps in their spectra at intermediate frequencies. The process with $\phi_1 = 0.3$, $\phi_2 = 0.4$ exhibits most of its variation at low frequencies, but also shows a smaller maximum at high frequencies. The spectrum for the process with $\phi_1 = 0.7$, $\phi_2 = -0.2$ resembles the red-noise spectra in Fig. 8.22, although with a broader and flatter low-frequency maximum.

Fig. 8.23 Theoretical spectral density functions for four AR(2) processes, computed using Eq. (8.83). The diversity of the forms of the spectra in this figure illustrates the flexibility of the AR(2) model. The autocorrelation functions for these autoregressions are shown as insets in Fig. 8.7.

Example 8.14. Smoothing a Sample Spectrum Using an Autoregressive Model

The equations for the theoretical spectra of autoregressive models can be useful in interpreting sample spectra from data series. The erratic sampling properties of the individual periodogram estimates can make it difficult to discern features of the true spectrum that underlies a particular sample spectrum. However, if a well-fitting time-domain model can be fit to the same data series, its theoretical spectrum can be used to guide the eye through the sample spectrum. Autoregressive models are sometimes fitted to time series for the sole purpose of obtaining smooth spectra. Chu and Katz (1989) show that the spectrum corresponding to a time-domain model fit using the SOI time series (Fig. 8.18) corresponds well to the spectrum computed directly from the data.

Consider the data series in Fig. 8.8c, which was generated according to the AR(2) process with $\phi_1 = 0.9$ and $\phi_2 = -0.6$. The sample spectrum for this particular batch of 50 points is shown as the solid curve in Fig. 8.24. Apparently the series exhibits pseudoperiodicities in the frequency range around $f = 0.12-0.16$, but sampling variability makes the interpretation somewhat difficult. Although the solid curve in Fig. 8.24 somewhat resembles the theoretical spectrum for this AR(2) model shown in Fig. 8.23, its nature might not be obvious from the empirical spectrum alone.

A fuller perspective on the spectrum in Fig. 8.24 is gained when the dashed curve is provided to guide the eye. This is the theoretical spectrum for the AR(2) model fitted to the same data points from which the empirical spectrum was computed. The first two sample autocorrelations for these data are $r_1 = 0.624$ and $r_2 = -0.019$, which are near the theoretical values that can be computed using Eqs. (8.33). Using Eqs. (8.29), the corresponding estimated autoregressive pa-

Fig. 8.24 The use of theoretical spectra for autoregressive models to guide the eye in interpreting sample spectra. The solid curve is the sample spectrum for the $n = 50$ data points shown in Fig. 8.8c, generated by the AR(2) process with $\phi_1 = 0.9$, $\phi_2 = -0.6$, and $\sigma_\epsilon^2 = 1.0$. A fuller perspective on this spectrum is provided by the dashed line, which is the theoretical spectrum of the AR(2) process fitted to this same series of 50 data points.

rameters are $\phi_1 = 1.04$ and $\phi_2 = -0.67$. The sample variance of the $n = 50$ data points is 1.69, which leads through Eq. (8.30) to the estimated white-noise variance $\sigma_\epsilon^2 = 1.10$. The resulting spectrum, according to Eq. (8.83), is plotted as the dashed curve. ¤

Exercises

8.1. Using the January 1987 precipitation data for Canandaigua in Table A.1
   (a) Fit a two-state, first-order Markov chain to represent daily precipitation occurrence.
   (b) Test whether this Markov model provides a significantly better representation of the data than does the assumption of independence.
   (c) Compare the theoretical stationary probability π_1 with the empirical relative frequency.
   (d) Graph the theoretical autocorrelation function for the first three lags.

8.2. Graph the autocorrelation functions up to five lags for
   (a) The AR(1) process with $\phi = 0.4$
   (b) The AR(2) process with $\phi_1 = 0.7$ and $\phi_2 = -0.7$

8.3. Computing sample lag correlations for a time series with $n = 100$ values, whose variance is 100, yields $r_1 = 0.80$, $r_2 = 0.60$, and $r_3 = 0.50$.
   (a) Use the Yule–Walker equations to fit AR(1), AR(2), and AR(3) models to the data. Assume that the sample size is large enough that Eq. (8.26) provides a good estimate for the white-noise variance.
   (b) Select the best autoregressive model for the series according to the BIC statistic.
   (c) Select the best autoregressive model for the series according to the AIC statistic.

8.4. Given that the mean of the time series in Exercise 8.3 is 50, use the fitted AR(2) model to forecast the future values of the time series x_1, x_2, and x_3, assuming that the current value is $x_0 = 76$ and the previous value is $x_{-1} = 65$.

8.5. The variance of a time series governed by the AR(1) model with $\phi = 0.8$ is 25. Compute the variances of the sampling distributions of averages of consecutive values of this time series, with lengths
   (a) $n = 5$
   (b) $n = 10$
   (c) $n = 50$

8.6. For the temperature data in Table 8.7
   (a) Calculate the first two harmonics.
   (b) Plot each of the two harmonics separately.
   (c) Plot the function representing the annual cycle defined by the first two

Table 8.7

Average Monthly Temperature Data for New Delhi, India

Month	J	F	M	A	M	J	J	A	S	O	N	D
Average temperature, °F	57	62	73	82	92	94	88	86	84	79	68	59

harmonics. Also include the original data points in this plot, and visually compare the goodness of fit.

(d) Represent these harmonics using a harmonic dial, with January 1 in the 12-o'clock position, and 1 cm representing 10° F.

8.7. Use the two-harmonic equation for the annual cycle from Exercise 8.6 to estimate the mean daily temperatures for

(a) April 10

(b) October 27

8.8. The amplitudes of the third, fourth, fifth, and sixth harmonics, respectively, of the data in Table 8.7 are 1.4907, 0.5773, 0.6311, and 0.0001° F.

(a) Plot a periodogram for this data. Explain what it shows.

(b) What proportion of the variation in the monthly average temperature data is described by the first two harmonics?

8.9. How many "tic-marks" for frequency are missing from the horizontal axis of Fig. 8.18?

8.10. Suppose that the minor peak in Fig. 8.18 at $f = 13/256 = 0.0508$ mo^{-1} resulted in part from aliasing.

(a) Compute a frequency that could have produced this spurious signal in the spectrum.

(b) How often would the underlying sea-level pressure data need to be recorded and processed in order to resolve this frequency explicitly?

8.11. Derive and plot the theoretical spectra for the two autoregressive processes in Exercise 8.2, assuming unit white-noise variance, and $n = 100$.

Chapter 9 Methods for Multivariate Data

9.1 Background

Much of the material in the first eight chapters of this book has pertained to analysis of univariate, or one-dimensional, data; that is, the analysis methods presented were oriented primarily toward scalar data values and their distributions. However, one finds in many practical situations that data sets are composed of vector observations. In this situation each data record consists of simultaneous observations of multiple quantities. Such data sets are known as *multivariate*. Examples of multivariate atmospheric data include simultaneous observations of different variables at one location, or an atmospheric field as represented by a set of grid-point values at a particular time.

Univariate methods can be, and are, applied to individual scalar elements of multivariate data observations. The distinguishing attribute of the methods of multivariate analysis is that both the joint behavior of the multiple simultaneous observations, as well as the variations of the individual data elements, are considered.

This chapter presents introductions to some of the multivariate methods that are used most commonly with atmospheric data. These include approaches to data reduction and structural simplification, summarization of multiple dependencies, and grouping and classification. Absent from this chapter is a presentation of multivariate hypothesis testing. Many multivariate testing methods require the often restrictive assumption of multivariate normality, and are covered extensively in textbooks devoted to multivariate statistics (e.g., Johnson and Wichern, 1982, or Morrison, 1967).

Multivariate methods are more difficult to understand and implement than univariate methods. Notationally, they require use of matrix algebra to make the presentation tractable. The elements of matrix algebra that are necessary to the subsequent material are presented in Section 9.2. The complexities of multivariate data and the methods that have been devised to deal with them dictate that all except the most trivial multivariate analyses will be implemented by computer. Enough detail is included here for readers comfortable with numerical methods to be able to implement the analyses themselves. However, many readers will use

"canned" packages for this purpose, and the material in this chapter should helpful in understanding what the programs are doing, and why.

9.2 Matrix Algebra Notation

The mathematics of dealing simultaneously with multiple variables and their mutual correlations is greatly simplified by use of a notation called *matrix algebra*, or *linear algebra*. A brief review of this subject, sufficient for the multivariate techniques described in the remainder of this chapter, is presented in this section. More complete introductions are readily available elsewhere (e.g., Lipschutz, 1968).

9.2.1 Vectors

The *vector* is a fundamental component of the notation of matrix algebra. It is essentially nothing more than an ordered list of scalar variables, or ordinary numbers. These numbers, also called *elements* of the vector, pertain to different aspects of a phenomenon or situation being measured or described. The number of elements, also called the vector's *dimensionality*, will depend on the situation at hand. A familiar meteorological example is the two-dimensional horizontal wind vector, whose two elements are the eastward windspeed u, and the northward windspeed v.

Here vectors will be denoted using boldface lowercase letters, and their individual elements will be distinguished by subscripts running from 1 to K. A vector with only $K = 1$ element is just an ordinary number, or scalar. Most commonly, vectors are *column vectors*, which means that their elements are arranged vertically. For example, the column vector $\mathbf{x}$ would consist of the elements x_1, x_2, $x_3, \ldots, x_K$; arranged as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix}. \tag{9.1}$$

These same elements can be arranged horizontally, as a *row vector*, through an operation called *transposing* the vector. The transpose operation is denoted by the superscript "T," so that one can write the vector $\mathbf{x}$ in Eq. (9.1) as the row vector

$$\mathbf{x}^{\mathrm{T}} = [x_1, x_2, x_3, \ldots, x_K], \tag{9.2}$$

which is pronounced "x-transpose." Using the transpose of a column vector is useful for notational consistency with certain matrix operations. It is also useful

for typographic purposes, as it allows one to write out a vector on a horizontal line of text.

Addition of two or more vectors with of the same dimension is straightforward. Vector addition is accomplished by adding the corresponding elements of the two vectors; for example

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \\ \vdots \\ x_K + y_K \end{bmatrix}. \tag{9.3}$$

Subtraction is accomplished analogously. Addition and subtraction of vectors with different dimensions is not defined.

Multiplying a vector by a scalar results in a new vector whose elements are simply the corresponding elements of the original vector multiplied by that scalar. For example, multiplying the vector $\mathbf{x}$ in Eq. (9.1) by a constant c yields

$$c\mathbf{x} = \begin{bmatrix} cx_1 \\ cx_2 \\ cx_3 \\ \vdots \\ cx_K \end{bmatrix}. \tag{9.4}$$

Two vectors of the same dimension can be multiplied using an operation called the *dot product*, or *inner product*. This operation consists of multiplying together each of the K like pairs of vector elements, and then summing these K products:

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = [x_1, x_2, x_3, \ldots, x_K]\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_K \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 + \cdots + x_Ky_K$$

$$= \sum_{k=1}^{K} x_k y_k. \tag{9.5}$$

This vector multiplication has been written as the product of a row vector and a column vector for consistency with the operation of matrix multiplication, presented in the following section. As will be seen, the dot product is in fact a special case of matrix multiplication. Equation (9.5) also shows that vector multiplication can also be expressed in *component form* using summation notation. Expanding vector and matrix operations in component form is useful if the calculation is to be programmed for a computer.

It is possible and often convenient to think of vectors geometrically. In this way

of thinking, a vector locates a point in the K-dimensional space whose Cartesian coordinates are each of the K elements of the vector. This is easiest to see in $K = 2$ dimensions, where the two-dimensional space is a plane. In this case x_1 is the horizontal coordinate, and x_2 is the vertical coordinate of the point $\mathbf{x}$. Similarly, one also sometimes thinks of a vector as an arrow "pointing" from the origin (located at $x_1 = x_2 = x_3 = \cdots = x_K = 0$) to the point $\mathbf{x}$. This way of thinking can be extended to higher dimensions, although for K greater than three dimensions the geometry can be difficult or impossible to visualize explicitly.

The length of a vector in its K-dimensional space is a scalar quantity that can be computed using the dot product, as

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \left[\sum_{k=1}^{K} x_k^2\right]^{1/2}. \tag{9.6}$$

Equation (9.6) is sometimes known as the *Euclidean norm* of the vector $\mathbf{x}$. In $K = 2$ dimensions it is easy to see that this length is simply an application of the Pythagorean theorem, or $\|\mathbf{x}\| = [x_1^2 + x_2^2]^{1/2}$. A common application of this equation is in the computation of the total horizontal windspeed from the horizontal velocity vector $\mathbf{v}^{\mathrm{T}} = [u, v]$, according to $v_{\mathrm{H}} = (u^2 + v^2)^{1/2}$. However, Eq. (9.6) generalizes to arbitrarily high K as well.

The angle θ between two vectors is also computed using the dot product, using

$$\theta = \cos^{-1}\left[\frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}\right]. \tag{9.7}$$

This relationship implies that two vectors are perpendicular if their product is zero, since $\cos[0] = 90°$. Mutually perpendicular vectors are also called *orthogonal.*

The magnitude of the projection (or "length of the shadow") of a vector $\mathbf{x}$ onto a vector $\mathbf{y}$ is also a function of the dot product, given by

$$L_{\mathbf{x},\mathbf{y}} = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{y}\|}. \tag{9.8}$$

The geometric interpretations of these three computations of length, angle, and projection are illustrated in Fig. 9.1, for the vectors $\mathbf{x}^{\mathrm{T}} = [1, 1]$ and $\mathbf{y}^{\mathrm{T}} = [2, 0.8]$. The length of $\mathbf{x}$ is simply $\|\mathbf{x}\| = (1^2 + 1^2)^{1/2} = \sqrt{2}$, and the length of $\mathbf{y}$ is $\|\mathbf{y}\| = (2^2 + 0.8^2)^{1/2} = 2.154$. Since the dot product of the two vectors is $\mathbf{x}^{\mathrm{T}}\mathbf{y} = 1 \cdot 2 + 1 \cdot 0.8 = 2.8$, the angle between them is $\theta = \cos^{-1}[2.8/(\sqrt{2} \cdot 2.154)] = 49°$, and the projection of $\mathbf{x}$ onto $\mathbf{y}$ is $2.8/2.154 = 1.302$.

9.2.2 Matrices

A matrix is a two-dimensional rectangular array of numbers having I rows and J columns. The dimension of a matrix is specified by the number of rows and columns. A matrix dimension is written "(I × J)", and pronounced "I by J." Matrices are denoted here by boldface uppercase letters surrounded by square brackets.

Fig. 9.1 Illustration of the concepts of vector length [Eq. (9.6)], the angle between two vectors [Eq. (9.7)], and the projection of one vector onto another [Eq. (9.8)]; for the two vectors $\mathbf{x}^T = [1, 1]$ and $\mathbf{y}^T = [2, 0.8]$.

The elements of a matrix are the individual variables or numerical values occupying the rows and columns. The matrix elements are identified notationally by two subscripts; the first of these identifies the row number and the second identifies the column number. For example, the (4×4) matrix **[A]** is a collection of the 16 individual elements $a_{i,j}$, arranged as

$$[\mathbf{A}] = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{bmatrix}. \tag{9.9}$$

A matrix with the same number of rows and columns, such as **[A]** in Eq. (9.9), is called a *square matrix*. The elements of a square matrix for which $i = j$ are arranged on the diagonal line from the upper left to the lower right corners, and are called *diagonal elements*. Correlation matrices **[R]** (Fig. 3.17) are square matrices having all 1's on the diagonal. A matrix for which $a_{i,j} = a_{j,i}$ for all values of i and j is called *symmetric*. Correlation matrices are also symmetric because the correlation between variable i and variable j is identical to the correlation between variable j and variable i. Another important square, symmetrical matrix is the *identity matrix* **[I]**, consisting of 1's on the diagonal and zeros everywhere else:

$$[\mathbf{I}] = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{9.10}$$

An identity matrix can be constructed for any (square) dimension. When the identity matrix appears in an equation, it can be assumed to be of appropriate dimension for the relevant matrix operations to be defined.

Square matrices are special cases for which the number of rows equals the number of columns, i.e., $I = J$. Not all matrices are square. For example, vectors are special cases of matrices where one of the dimensions is 1. A K-dimensional column vector is a (K $\times$ 1) matrix, and a K-dimensional row vector is a (1 $\times$ K) matrix.

The transpose operation is defined for any matrix, including the special case of vectors. The transpose of a matrix is obtained in general by exchanging row and column indices, not by a 90° rotation as might have been anticipated from Eq. (9.2). For example, the relationship between the (3 $\times$ 4) matrix **[B]** and its transpose, the (4 $\times$ 3) matrix $\mathbf{[B]}^{\mathrm{T}}$ is illustrated by comparing

$$\mathbf{[B]} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & b_{1,4} \\ b_{2,1} & b_{2,2} & b_{2,3} & b_{2,4} \\ b_{3,1} & b_{3,2} & b_{3,3} & b_{3,4} \end{bmatrix} \tag{9.11a}$$

and

$$\mathbf{[B]}^{\mathrm{T}} = \begin{bmatrix} b_{1,1} & b_{2,1} & b_{3,1} \\ b_{1,2} & b_{2,2} & b_{3,2} \\ b_{1,3} & b_{2,3} & b_{3,3} \\ b_{1,4} & b_{2,4} & b_{3,4} \end{bmatrix}. \tag{9.11b}$$

If the matrix **[A]** is symmetric, then $\mathbf{[A]}^{\mathrm{T}} = \mathbf{[A]}$.

Multiplication of a matrix by a scalar is the same as for vectors, and is accomplished by multiplying each element of the matrix by the scalar:

$$c\mathbf{[D]} = c\begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} = \begin{bmatrix} cd_{1,1} & cd_{1,2} \\ cd_{2,1} & cd_{2,2} \end{bmatrix}. \tag{9.12}$$

Similarly, matrix addition and subtraction are defined only for matrices of identical dimension, and are accomplished by performing these operations on the elements in the corresponding row and column positions. For example, the sum of two (2 $\times$ 2) matrices would be computed as

$$\begin{aligned} \mathbf{[D]} + \mathbf{[E]} &= \begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} + \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} d_{1,1} + e_{1,1} & d_{1,2} + e_{1,2} \\ d_{2,1} + e_{2,1} & d_{2,2} + e_{2,2} \end{bmatrix}. \end{aligned} \tag{9.13}$$

Matrix multiplication is defined between two matrices if the number of columns in the left matrix is equal to the number of rows in the right matrix. Thus, not only is matrix multiplication not commutative (i.e., **[A][B]** $\neq$ **[B][A]**), but multiplica-

tion of two matrices in reverse order is not even defined unless both are square and of the same dimension. The product of a matrix multiplication is another matrix, the row dimension of which is the same as the row dimension of the left matrix, and the column dimension of which is the same as the column dimension of the right matrix. In other words, multiplying a $(I \times J)$ matrix **[A]** and a $(J \times K)$ matrix **[B]** yields a $(I \times K)$ matrix **[C]**. In effect, the "middle" dimension J is multiplied out.

Consider the case where $I = 2$, $J = 3$, and $K = 2$. In terms of the individual matrix elements, the matrix multiplication **[A][B]** = **[C]** expands to

$$\underset{(2 \times 3)}{\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}} \underset{(3 \times 2)}{\begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \\ b_{3,1} & b_{3,2} \end{bmatrix}} = \underset{(2 \times 2)}{\begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix}}, \tag{9.14a}$$

where

$$\begin{aligned} [\mathbf{C}] &= \begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} + a_{1,3}b_{3,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} + a_{1,3}b_{1,3} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} + a_{2,3}b_{3,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} + a_{2,3}b_{3,2} \end{bmatrix}. \end{aligned} \tag{9.14b}$$

The individual components of **[C]** as written out in Eq. (9.14b) may look confusing at first exposure. In understanding matrix multiplication, it is helpful to realize that each element of the product matrix **[C]** is simply the dot product, as defined in Eq. (9.5), of one of the rows in the left matrix **[A]** and one of the columns in the right matrix **[B]**. In particular, the number occupying the ith row and kth column of the matrix **[C]** is exactly the dot product between the row *vector* making up the ith row of **[A]** and the column *vector* constituting the kth column of **[B]**. Equivalently, matrix multiplication can be written in terms of the individual matrix elements using summation notation,

$$c_{i,k} = \sum_{j=1}^{J} a_{i,j}b_{j,k}; \qquad i = 1, \ldots, I; \quad k = 1, \ldots, K. \tag{9.15}$$

The analog of arithmetic division exists for square matrices that have a property known as *full rank*, or *nonsingularity*. This condition can be interpreted to mean that the matrix does not contain redundant information, in the sense that none of the rows can be constructed from linear combinations of the other rows. Considering each row of the nonsingular matrix as a vector, it is impossible to construct vector sums of rows multiplied by scalar constants, that equal any one of the other rows. These same conditions applied to the columns also imply that the matrix is nonsingular.

Nonsingular square matrices are *invertable*; for example, a matrix **[A]** is invertable if another matrix **[B]** exists such that

$$[\mathbf{A}][\mathbf{B}] = [\mathbf{B}][\mathbf{A}] = [\mathbf{I}]. \tag{9.16}$$

It is then said that **[B]** is the inverse of **[A]**, or $[\mathbf{B}] = [\mathbf{A}]^{-1}$; and that **[A]** is the inverse of **[B]**, or $[\mathbf{A}] = [\mathbf{B}]^{-1}$. Inverses of (2×2) matrices are easy to compute by hand, using

$$[\mathbf{A}]^{-1} = \frac{1}{(a_{1,1}a_{2,2}) - (a_{2,1}a_{1,2})} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix}. \tag{9.17}$$

This matrix is pronounced "A inverse." Explicit formulas for inverting matrices of higher dimension also exist, but quickly become very cumbersome as the dimensions get larger. Computer algorithms for inverting matrices are widely available, and as a consequence matrices with dimension higher than two or three are rarely inverted by hand. An important exception is the inverse of a diagonal matrix, which is simply another diagonal matrix whose nonzero elements are the reciprocals of the diagonal elements of the original matrix.

Table 9.1 lists some properties of arithmetic operations with matrices that have not been specifically mentioned in the foregoing.

Example 9.1. Computation of the Correlation Matrix

The correlation matrix **[R]** was introduced in Chapter 3 as a device for compactly representing the mutual correlations among K variables. A generalized correlation matrix is shown in Fig. 3.17, and the correlation matrix for the January 1987 data in Appendix A.1 (with the unit diagonal elements and the symmetry implicit) is shown in Table 3.3. The computation of these correlations given in Eq. (3.18) can also be expressed in notation of matrix algebra.

The computation begins with the $(n \times K)$ matrix **[X]** of data values whose correlations are to be computed. Each row of this matrix is a vector, consisting of

Table 9.1

Some Elementary Properties of Arithmetic Operations with Matrices

Multiplication by the identity matrix	$[\mathbf{A}][\mathbf{I}] = [\mathbf{I}][\mathbf{A}] = [\mathbf{A}]$
Distributive multiplication by a scalar	$c([\mathbf{A}][\mathbf{B}]) = (c[\mathbf{A}])[\mathbf{B}]$
Distributive matrix multiplication	$[\mathbf{A}]([\mathbf{B}] + [\mathbf{C}]) = [\mathbf{A}][\mathbf{B}] + [\mathbf{A}][\mathbf{C}]$
	$([\mathbf{A}] + [\mathbf{B}])[\mathbf{C}] = [\mathbf{A}][\mathbf{C}] + [\mathbf{B}][\mathbf{C}]$
Associative matrix multiplication	$[\mathbf{A}]([\mathbf{B}][\mathbf{C}]) = ([\mathbf{A}][\mathbf{B}])[\mathbf{C}]$
Inverse of a matrix product	$([\mathbf{A}][\mathbf{B}])^{-1} = [\mathbf{B}]^{-1}[\mathbf{A}]^{-1}$
Transpose of a matrix product	$([\mathbf{A}][\mathbf{B}])^{\mathrm{T}} = [\mathbf{B}]^{\mathrm{T}}[\mathbf{A}]^{\mathrm{T}}$
Combining matrix transpose and inverse	$([\mathbf{A}]^{-1})^{\mathrm{T}} = ([\mathbf{A}]^{\mathrm{T}})^{-1}$

one observation each of K variables. The number of these rows is the same as the sample size, n, since each of these row vectors is a sample point (in K dimensions). Thus, **[X]** is essentially just an ordinary data table such as Table A.1. In Table A.1 there are $K = 6$ variables (excluding the column containing the dates), each simultaneously observed on $n = 31$ occasions. An individual data element $x_{i,k}$ is the ith observation of the kth variable. For example, in Table A.1, $x_{4,6}$ would be the Canandaigua minimum temperature (19° F) observed on January 4.

Define the $(n \times n)$ matrix **[1]**, whose elements are all equal to 1. The $(n \times K)$ matrix of anomalies (in the meteorological sense of variables with their means subtracted), or *centered data* **[X′]** is then

$$[\mathbf{X}'] = [\mathbf{X}] - \frac{1}{n}[\mathbf{1}][\mathbf{X}]. \tag{9.18}$$

The second term in Eq. (9.18) is a $(n \times K)$ matrix containing the sample means. Its n rows are all the same, and each consists of the K sample means in the same order as the corresponding variables appear in each row of **[X]**.

Multiplying **[X′]** by the transpose of itself, and dividing by $n - 1$, yields an important matrix called the *variance–covariance matrix*, or sometimes simply the *variance matrix*,

$$[\mathbf{S}] = \frac{1}{n-1}[\mathbf{X}']^{\mathrm{T}}[\mathbf{X}']. \tag{9.19}$$

The variance–covariance matrix **[S]** is a symmetric $(K \times K)$ matrix whose diagonal elements are the sample variances of the K variables, and whose other elements are the covariances among the K variables. For example, $s_{1,1}$ is the sample variance of the first variable, and $s_{1,2}$ is the sample covariance between the first and the second variables. Because variances are so often expressed in terms of variance–covariance matrices, it is fairly common to see the notation $s_{i,i}$ for the sample variance, which is equivalent to the more common and familiar s_i^2. The operation in Eq. (9.19) corresponds to the summation in the numerator of Eq. (3.18).

The variance–covariance matrix is also known as the *dispersion matrix*, because it describes how the observations are dispersed around their (vector) mean in the K-dimensional space defined by the K variables. The diagonal elements are the individual variances, which index the degree to which the data are spread out in directions parallel to the K axes, and the covariances in the off-diagonal positions describe the extent to which the cloud of data points is oriented at angles to these axes. The matrix **[S]** is the sample estimate of the population dispersion matrix $[\mathbf{\Sigma}]$, which appears in the probability density function for the multivariate normal distribution [Eq. (4.50)].

Now define the $(K \times K)$ diagonal matrix **[D]**, whose diagonal elements are the sample standard deviations of the K variables; That is, **[D]** consists of all zeros

except for the diagonal elements, whose values are the square roots of the corresponding elements of **[S]**. That is, $d_{k,k} = \sqrt{s_{k,k}}$, $k = 1, \ldots, K$. The correlation matrix can then be computed from the variance–covariance matrix using

$$[\mathbf{R}] = [\mathbf{D}]^{-1}[\mathbf{S}][\mathbf{D}]^{-1}. \tag{9.20}$$

Since **[D]** is diagonal, its inverse is the diagonal matrix whose elements are the reciprocals of the sample standard deviations on the diagonal of **[D]**. The matrix multiplication in Eq. (9.20) corresponds to division by the standard deviations in Eq. (3.18).

Note that the correlation matrix **[R]** is equivalently the variance–covariance matrix of the standardized variables (or *standardized anomalies*) z_k [Eq. (3.16)]. That is, dividing the anomalies x'_k by their standard deviations $\sqrt{s_{k,k}}$ nondimensionalizes the variables, and results in their having unit variance (1's on the diagonal of **[R]**) and covariances equal to their correlations. In matrix notation this can be seen by substituting Eq. (9.19) into Eq. (9.20) to yield

$$\begin{aligned}[\mathbf{R}] &= \frac{1}{n-1}[\mathbf{D}]^{-1}[\mathbf{X}']^{\mathrm{T}}[\mathbf{X}'][\mathbf{D}]^{-1} \\ &= \frac{1}{n-1}[\mathbf{Z}]^{\mathrm{T}}[\mathbf{Z}],\end{aligned} \tag{9.21}$$

where **[Z]** is the $(n \times K)$ matrix whose rows are the vectors of standardized variables **z**, analogously to the matrix **[X′]** of the anomalies. The first line of Eq. (9.21) converts the matrix **[X′]** to the matrix **[Z]** by dividing each element by its standard deviation, $d_{k,k}$. Comparison of Eqs. (9.21) and (9.19) shows that **[R]** is, indeed, the variance–covariance matrix for the standardized variables **z**. ¤

Example 9.2. Multiple Regression Expressed in Matrix Notation

The discussion of multiple linear regression in Chapter 6 indicated that the relevant mathematics are most easily expressed and solved using matrix algebra. In this notation, the expression for the predictand y as a function of the predictor variables x_i [Eq. (6.22)] becomes

$$\mathbf{y} = [\mathbf{X}]\mathbf{b}, \tag{9.22a}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,K} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}. \tag{9.22b}$$

Here $\mathbf{y}$ is a $(n \times 1)$ matrix (i.e., a vector) of the n observations of the predictand, $[\mathbf{X}]$ is a $(n \times K + 1)$ data matrix containing the values of the predictor variables, and $\mathbf{b}^{\mathrm{T}} = [b_0, b_1, b_2, \ldots, b_K]$ is a $(K + 1 \times 1)$ vector of the regression parameters. The data matrix in the regression context is similar to that used in Example 9.1, except that it has $K + 1$ rather than K columns. This extra column is the leftmost column of $[\mathbf{X}]$, and consists entirely of 1's. Thus, Eq. (9.22) is a vector equation, with dimension $(n \times 1)$ on each side. It is actually n repetitions of Eq. (6.22), once each for the n data records.

The "normal equations" [presented in Eq. (6.6) for the simple case of $K = 1$] are obtained by multiplying each side of Eq. (9.22) by $[\mathbf{X}]^{\mathrm{T}}$:

$$[\mathbf{X}]^{\mathrm{T}}\mathbf{y} = [\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]\mathbf{b}, \tag{9.23a}$$

or

$$\begin{bmatrix} \Sigma\, y \\ \Sigma\, x_1 y \\ \Sigma\, x_2 y \\ \vdots \\ \Sigma\, x_K y \end{bmatrix} = \begin{bmatrix} n & \Sigma\, x_1 & \Sigma\, x_2 & \cdots & \Sigma\, x_K \\ \Sigma\, x_1 & \Sigma\, x_1^2 & \Sigma\, x_1 x_2 & \cdots & \Sigma\, x_1 x_K \\ \Sigma\, x_2 & \Sigma\, x_2 x_1 & \Sigma\, x_2^2 & \cdots & \Sigma\, x_2 x_K \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma\, x_K & \Sigma\, x_K x_1 & \Sigma\, x_K x_2 & \cdots & \Sigma\, x_K^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}, \tag{9.23b}$$

where all the summations are over the n data points. Here the $[\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]$ matrix has dimension $(K + 1 \times K + 1)$. Each side of Eq. (9.23) has dimension $(K + 1 \times 1)$, and this equation actually represents $K + 1$ simultaneous equations containing the $K + 1$ unknown regression coefficients. Matrix algebra is very commonly used to solve sets of simultaneous linear equations such as these. One way to obtain the solution is to multiply both sides of Eq. (9.23) by the inverse of the $[\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]$ matrix. This operation is analogous to dividing both sides by this quantity, and yields

$$\begin{aligned} ([\mathbf{X}]^{\mathrm{T}}[\mathbf{X}])^{-1}[\mathbf{X}]^{\mathrm{T}}\mathbf{y} &= ([\mathbf{X}]^{\mathrm{T}}[\mathbf{X}])^{-1}[\mathbf{X}]^{\mathrm{T}}[\mathbf{X}]\mathbf{b} \\ &= [\mathbf{I}]\mathbf{b} \\ &= \mathbf{b}, \end{aligned} \tag{9.24}$$

which is the vector of regression parameters. ¤

9.2.3 Eigenvalues and Eigenvectors of a Square Matrix

An eigenvalue λ, and an eigenvector, $\mathbf{e}$, of a square matrix $[\mathbf{B}]$ are a scalar and nonzero vector, respectively, satisfying the equation

$$[\mathbf{B}]\mathbf{e} = \lambda\mathbf{e}, \tag{9.25a}$$

or equivalently

$$([\mathbf{B}] - \lambda[\mathbf{I}])\mathbf{e} = \mathbf{0}, \tag{9.25b}$$

where $\mathbf{0}$ is a vector consisting entirely of zeros. For every eigenvalue and eigen-

vector pair that can be found to satisfy Eq. (9.25), any scalar multiple of the eigenvector, $c\mathbf{e}$, will also satisfy the equation together with that eigenvalue. Consequently, it is usual to require that the eigenvalues have unit length:

$$\|\mathbf{e}\| = 1. \tag{9.26}$$

This restriction removes the ambiguity only up to a change in sign, since if a vector $\mathbf{e}$ satisfies Eq. (9.25), its negative, $-\mathbf{e}$ will, also.

If $[\mathbf{B}]$ is nonsingular there will be $M = K$ nonzero eigenvalue/eigenvector pairs λ_m and $\mathbf{e}_m$, where K is the number of rows and columns in $[\mathbf{B}]$. Each eigenvector will be dimensioned $(M \times 1)$. Synonymous terminology that is sometimes also used for eigenvalues and eigenvectors includes *characteristic values* and *characteristic vectors, latent values* and *latent vectors*, and *proper values* and *proper vectors.*

For many statistical applications, eigenvalues and eigenvectors are calculated for real (not containing complex or imaginary numbers), symmetric matrices. Eigenvalues and eigenvectors of such matrices have a number of important and remarkable properties, some of which also hold for the eigenvalues and eigenvectors of square matrices more generally. The first of these properties is that the eigenvectors of symmetric matrices are orthogonal. That is, their dot products with each other are zero. Furthermore, since each eigenvector is defined to have unit length, the dot product of any eigenvector with itself is one:

$$\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \tag{9.27}$$

Orthogonal vectors of unit length are said to be *orthonormal.* (This terminology has nothing to do with the Gaussian, or "normal" distribution.) The orthonormality property is analogous to Eq. (8.71), expressing the orthogonality of the sine and cosine functions. Often the $(M \times M)$ matrix $[\mathbf{E}]$ is formed, the M columns of which are the eigenvectors $\mathbf{e}_k$. Because of the orthogonality and unit length of the eigenvectors,

$$[\mathbf{E}]^{\mathrm{T}}[\mathbf{E}] = [\mathbf{I}], \tag{9.28a}$$

which implies

$$[\mathbf{E}]^{-1} = [\mathbf{E}]^{\mathrm{T}}. \tag{9.28b}$$

Matrices exhibiting these properties are said to be *orthogonal.*

The matrix of eigenvectors $[\mathbf{E}]$ has the property that it *diagonalizes* the original matrix $[\mathbf{B}]$ from which the eigenvectors and eigenvalues were calculated. This diagonalization is expressed mathematically as

$$[\mathbf{E}]^{-1}[\mathbf{B}][\mathbf{E}] = [\Lambda] = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_M \end{bmatrix}. \tag{9.29}$$

Multiplication of **[B]** on the left by $[\mathbf{E}]^{-1}$ and on the right by **[E]** produces the diagonal matrix $[\Lambda]$, the elements of which are the eigenvalues λ_m corresponding to the eigenvectors $\mathbf{e}_m$ in each of the K columns of **[E]**. The eigenvalues themselves have the property that their sum is equal to the sum of the diagonal elements of the original matrix **[B]**. That is

$$\sum_{m=1}^{M} \lambda_m = \sum_{k=1}^{K} b_{k,k}. \tag{9.30}$$

The extraction of eigenvalue/eigenvector pairs from matrices is a computationally demanding task, particularly as the dimensionality of the problem increases. It is possible but very tedious to do the computations by hand if $K = M = 2$, 3, or 4. This calculation first requires solving a Kth-order polynomial for the K eigenvalues, and then solving K sets of K simultaneous equations to obtain the eigenvectors. In general, however, different algorithms are used. Computer routines for calculating numerical approximations to eigenvalues and eigenvectors are widely available, particularly for the real and symmetrical matrices that are most often of interest in statistical work.

Example 9.3. Eigenvalues and Eigenvectors of a (2 × 2) Symmetric Matrix

The symmetric matrix

$$[\mathbf{B}] = \begin{bmatrix} 185.47 & 110.84 \\ 110.84 & 77.58 \end{bmatrix} \tag{9.31}$$

has as its eigenvalues $\lambda_1 = 254.76$ and $\lambda_2 = 8.29$, with corresponding eigenvectors $\mathbf{e}_1^{\mathrm{T}} = [0.848, 0.530]$ and $\mathbf{e}_2^{\mathrm{T}} = [-0.530, 0.848]$. It is easily verified that both eigenvalues are of unit length. Their dot product is zero, which indicates that the two vectors are perpendicular, or orthogonal.

The matrix of eigenvectors is therefore

$$[\mathbf{E}] = \begin{bmatrix} 0.848 & -0.530 \\ 0.530 & 0.848 \end{bmatrix}, \tag{9.32}$$

and this matrix of eigenvectors diagonalizes the original matrix **[B]** according to

$$\begin{aligned} [\mathbf{E}]^{-1}[\mathbf{B}][\mathbf{E}] &= \begin{bmatrix} 0.848 & 0.530 \\ -0.530 & 0.848 \end{bmatrix} \begin{bmatrix} 185.47 & 110.84 \\ 110.84 & 77.58 \end{bmatrix} \begin{bmatrix} 0.848 & -0.530 \\ 0.530 & 0.848 \end{bmatrix} \\ &= \begin{bmatrix} 254.76 & 0 \\ 0 & 8.29 \end{bmatrix} = [\Lambda]. \end{aligned} \tag{9.33}$$

Because of the orthogonality of the eigenvectors, the inverse of **[E]** is equal to its transpose as indicated in Eq. (9.33). Finally, the sum of the eigenvectors, 254.76 +

8.29 = 263.05, equals the sum of the diagonal elements of the original **[B]** matrix, 185.47 + 77.58 = 263.05. ¤

Example 9.4. Confidence Ellipses for the Bivariate Normal Distribution

Recall from Chapter 4 that curves of constant height on the probability density function (PDF) for the bivariate normal distribution [Eq. (4.45)] are elliptical, and centered on the vector mean. The equations for these ellipses are of interest because the volumes under the density function enclosed by them correspond to particular probabilities. The ellipses are sometimes known as *confidence ellipses.*

The directions of the major and minor axes of these ellipses are given by the eigenvectors of the variance–covariance matrix of the two variables. The degrees to which these ellipses are stretched in these two directions are proportional to the square roots of the respective eigenvalues. These concepts are illustrated for the minimum temperature data in Table A.1, the variance–covariance matrix for which is given in Eq. (9.31). The direction of the major axis is given by the first eigenvector, $\mathbf{e}_1^T = [0.848, 0.530]$, whose angle with the x_1 (Ithaca temperature) axis is 32° [from Eq. (9.7)]. The ellipses are stretched much more in the $\mathbf{e}_1$ than in the $\mathbf{e}_2$ direction because the first eigenvalue is so much larger than the second.

The constants multiplying the square roots of the eigenvalues depend on the amount of probability each ellipse encloses. These constants can be obtained from cumulative probabilities of the χ^2 distribution. This distribution is a special case of the gamma distribution, for which cumulative probabilities can be obtained using Table B.2. For the bivariate normal distribution, the distance from the center to the curve of the ellipse, along the direction of the axes, is given by the product $[\lambda_m \chi_2^2(F)]^{1/2}$, where $\chi_2^2(F)$ is the value of the chi-square variable with $\nu = 2$ degrees of freedom corresponding to a cumulative probability F. Equivalently, this is the value from the gamma distribution with $\alpha = 1$ and $\beta = 2$ at the same cumulative probability. In Fig. 9.2, the 50% and 90% confidence ellipses are derived from $F = 0.5$ and $F = 0.9$. Doubling the values in Table B.2 (because $\beta = 2$), these distances are $[1.39\lambda_m]^{1/2}$ and $[4.60\lambda_m]^{1/2}$, respectively.

These ideas generalize directly for the multivariate normal distribution [Eq. (4.50)]. Here, there are $K > 2$ variables in the data vector $\mathbf{x}$, and the axes of the confidence (hyper) ellipses are oriented in the directions of the eigenvectors of the $(K \times K)$ variance-covariance matrix. The equation defining these ellipses is

$$(\mathbf{x} - \boldsymbol{\mu})^T[\mathbf{S}]^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \chi_K^2(F) = F_\Gamma^{-1}\left(\frac{K}{2}, 2\right), \tag{9.34}$$

which holds for the bivariate ($K = 2$) normal distribution as well. Here $F_\Gamma(\alpha, \beta)$ denotes the cumulative gamma distribution function. In general, the constant by which the square roots of the eigenvalues must be multiplied relates to the

Fig. 9.2 (50%) Inner and (90%) outer confidence ellipses for the bivariate normal distribution representing the minimum temperature data in Table A.1. The major and minor axes of the ellipses are oriented in the directions of the two eigenvectors of the variance–covariance matrix in Eq. (9.31), and are stretched in these directions by an amount proportional to the square roots of the respective eigenvalues. The ellipses are centered at the point defined by the two mean values.

χ^2 distribution with $\nu = K$ degrees of freedom, or the gamma distribution with $\alpha = K/2$. ¤

9.3 Principal-Component (EOF) Analysis

Possibly the most widely used multivariate statistical technique in the atmospheric sciences is *principal-component analysis* (PCA). The technique became popular for analysis of atmospheric data following the paper by Lorenz (1956), who called the technique *empirical orthogonal function* (EOF) analysis. Both names are commonly used, and refer to the same set of procedures. Sometimes the method is incorrectly referred to as *factor analysis*, which is a related but distinct multivariate statistical method.

The purpose of PCA is to reduce a data set containing a large number of variables to a data set containing fewer (hopefully many fewer) new variables, but that nevertheless represent a large fraction of the variability contained in the original data. That is, given multiple observations of a $(K \times 1)$ data vector $\mathbf{x}$, one hopes to find $(M^* \times 1)$ vectors $\mathbf{u}$ whose elements are functions of the elements of the $\mathbf{x}$ values, that contain most of the information in the original collection of $\mathbf{x}$ values,

and whose dimensionality $M^* << K$. This goal can be achieved if there are substantial correlations among the variables contained in **x**, in which case **x** contains redundant information. The elements of these new vectors **u** are called the *principal components*. In addition to constituting a compact representation of the original data **x**, the new variables comprising **u** exhibit the very desirable attribute of being mutually uncorrelated.

Usually the data for atmospheric, and other geophysical, fields exhibit many large correlations among the variables x_k, and a PCA results in a much more compact representation of their variations. Beyond mere data compression, however, a PCA can be a very useful tool for exploring large multivariate data sets, especially those consisting of geophysical fields. Here PCA has the potential for yielding substantial insights into both the spatial and temporal variations exhibited by the field or fields being analyzed.

This section is intended to provide a basic introduction to what has become a very large subject. Book-length treatments of PCA are given in Preisendorfer (1988), which is oriented specifically toward geophysical data; and in Jolliffe (1986), which describes PCA more generally. In addition, most textbooks on multivariate statistical analysis contain chapters on PCA.

9.3.1 Basics of PCA

At root, a PCA is based on analysis of the variance–covariance matrix **[S]** [Eq. (9.19)]. This matrix contains the sample variances of the K elements of the data vector **x** on its diagonal, and the covariances among these variables in the off-diagonal positions. Thus, **[S]** contains a wealth of information about the nature of the multiple variations exhibited by x. The variance–covariance matrix is related to the correlation matrix **[R]** through Eq. (9.20), which shows that **[R]** is a non-dimensionalized version of **[S]**. That is, **[R]** is produced by dividing each element of **[S]** by the standard deviations $\sqrt{s_{i,i}}$ and $\sqrt{s_{j,j}}$ of the variables occupying the ith row and jth column.

PCAs are conducted on *centered data*, or *anomalies* in the terminology of the atmospheric sciences. An anomaly x'_k is obtained simply by subtracting the sample mean of the x_k values. Similarly a vector of anomalies $\mathbf{x}'$ is computed simply as

$$\mathbf{x}' = \mathbf{x} - \overline{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix} - \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \overline{x}_3 \\ \vdots \\ \overline{x}_K \end{bmatrix}. \tag{9.35}$$

Anomalies differ from the original data only with respect to their mean values, which are all zero. Anomalies retain the same physical dimensions as the origi-

nal variables, and exhibit the same variances and mutual correlations. Indeed, Eq. (9.19) shows that the variance–covariance matrices of $\mathbf{x}$ and $\mathbf{x}'$ are identical.

The new variables, u_m, that will account successively for the maximum amount of the joint variability of $\mathbf{x}'$ (and therefore also of $\mathbf{x}$) are found using the eigenvectors of [S]. In particular, the mth principal component, u_m is obtained as the projection of the data vector $\mathbf{x}'$ onto the mth eigenvector, $\mathbf{e}_m$,

$$u_m = \mathbf{e}_m^{\mathrm{T}}\mathbf{x}' = \sum_{k=1}^{K} e_{km}x'_k, \qquad m = 1, \ldots, M. \tag{9.36}$$

Notice that each of the M eigenvectors contains one element pertaining to each of the K variables, x_k. Similarly, the mth principal component in Eq. (9.36) is computed from a particular set of observations of the K variables x_k. That is, each of the M principal components is a sort of weighted average of the x'_k values. Although the weights (the e_{km} values) do not sum to 1, their squares do because of the scaling convention $\|\mathbf{e}_m\| = 1$. If the data sample consists of n observations (and therefore of n data records, or n rows in the data matrix [X]), there will be n values for each of the principal components, or new variables, u_m.

Geometrically, the first eigenvector, $\mathbf{e}_1$, points to the direction (in the K-dimensional space of $\mathbf{x}'$) in which the data vectors jointly exhibit the most variability. This first eigenvector is the one associated with the largest eigenvalue, λ_1. The second eigenvector $\mathbf{e}_2$, associated with the second-largest eigenvalue λ_2, is constrained to be perpendicular to $\mathbf{e}_1$ [Eq. (9.27)], but subject to this constraint it will align in the direction that the $\mathbf{x}'$ vectors exhibit their next strongest variations. Subsequent eigenvectors $\mathbf{e}_m$, $m = 3, 4, \ldots, M$, are similarly numbered in decreasing magnitudes of their associated eigenvalues, and will in turn be perpendicular to all the previous eigenvectors. Subject to this constraint these eigenvectors will continue to locate directions in which the original data jointly exhibit maximum variability.

Put another way, the eigenvectors define a new coordinate system in which to view the data. This coordinate system is oriented such that each new axis is aligned along the direction of the maximum joint variability of the data, consistent with that axis being orthogonal to the preceding ones. These axes will turn out to be different for different data sets, because they are extracted from the sample variance-covariance matrix [S] particular to a given data set. That is, they are orthogonal functions, but are defined empirically according to the particular data set at hand. This observation is the basis for the eigenvectors being known in this context as EOFs. The implied distinction is with theoretical orthogonal functions, such as Fourier harmonics or Tschebyschev polynomials, which can also be used to define alternative coordinate systems in which to view a data set.

It is a remarkable property of the principal components that they are uncorrelated. That is, the correlation matrix for the new variables u_m is simply [I]. This implies that the covariances between pairs of the u_m values are also zero, so that

the corresponding variance–covariance matrix is diagonal. In fact, the variance–covariance matrix for the principal components is obtained by the diagonalization of [S] [Eq. (9.29)], and is thus simply the diagonal matrix $[\Lambda]$ of the eigenvalues of [S]. That is, the variance of the mth principal component u_m is the mth eigenvalue λ_m. Equation (9.30) then implies that the total variation exhibited by the x'_k values is completely represented in (or "accounted for" by) the u_m values, in the sense that the sum of the variances of the centered data $\mathbf{x}'$ (and therefore also of the uncentered variables $\mathbf{x}$) is equal to the sum of the variances of the new variables $\mathbf{u}$.

The expression for the $(M \times 1)$ vector of principal components $\mathbf{u}$, corresponding to Eq. (9.36), is

$$\mathbf{u} = [\mathbf{E}]^{\mathrm{T}}\mathbf{x}', \tag{9.37}$$

where, as before, [E] is the $(M \times K)$ square matrix whose columns are the eigenvectors of [S]. The matrix [E] is said to transform the data vector $\mathbf{x}'$ to the vector of new variables $\mathbf{u}$. This equation is sometimes called the *analysis formula* for $\mathbf{x}'$, expressing that the data can be analyzed in terms of the principal components. Conversely, the original data can be recovered from the principal components through the reverse transformation

$$\mathbf{x}' = [\mathbf{E}]\mathbf{u}, \tag{9.38a}$$

or

$$x'_k = \sum_{m=1}^{M} e_{km}u_m, \qquad k = 1, \ldots, K, \tag{9.38b}$$

which is obtained from Eq. (9.37) by multiplying on the left by [E] and using the orthogonality property of this matrix [Eq. (9.28)]. Equation (9.38) is sometimes called the *synthesis formula*, as it expresses the fact that the original centered data can be reconstructed from the principal components. The original uncentered data $\mathbf{x}$ can easily be obtained by adding back the vector of sample means, that is, by reversing Eq. (9.35).

Example 9.5. PCA in Two Dimensions

The basics of PCA are most easily appreciated in a simple example where the geometry can be visualized. If $K = 2$, the space of the data is two-dimensional, and can be graphed on a page. Figure 9.3 shows a scatterplot of centered January 1987 Ithaca minimum temperatures (x'_1) and Canandaigua minimum temperatures (x'_2) from Table A.1. This is the same scatterplot as appears in the middle of the bottom row of Fig. 3.18. Both variables are centered at zero. It is also apparent that the Ithaca temperatures are more variable than the Canandaigua temperatures, with the two standard deviations being $\sqrt{s_{1,1}} = 13.62°\mathrm{F}$ and $\sqrt{s_{2,2}} = 8.81°\mathrm{F}$, respectively. The two variables are clearly strongly correlated, and have a Pearson

Fig. 9.3 Scatterplot of January 1987 Ithaca and Canandaigua minimum temperatures (converted to anomalies, or "centered"), illustrating the geometry of PCA in two dimensions. The eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of the variance–covariance matrix **[S]** for these two variables, as computed in Example 9.3, have been plotted with exaggerated lengths for clarity. The data stretch out in the direction of $\mathbf{e}_1$ to the extent that 96.8% of the joint variance of these two variables occurs along this axis. The coordinates u_1 and u_2, corresponding to the data point $\mathbf{x}'^{\mathrm{T}} = [16.0, 17.8]$, recorded on January 15 and indicated by the large square symbol, are shown by lengths in the directions of the new coordinate system defined by the eigenvectors. That is, the coordinations $\mathbf{u}^{\mathrm{T}} = [23.0, 6.6]$ locate the same data point as $\mathbf{x}'^{\mathrm{T}} = [16.0, 17.8]$.

correlation of +0.924 (Table 3.3). The covariance matrix **[S]** for these two variables is given as **[B]** in Eq. (9.31). The two eigenvectors of this matrix are $\mathbf{e}_1^{\mathrm{T}} = [0.848, 0.530]$ and $\mathbf{e}_2^{\mathrm{T}} = [-0.530, 0.848]$, so that the eigenvector matrix **[E]** is as shown in Eq. (9.32). The corresponding eigenvalues are $\lambda_1 = 254.76$ and $\lambda_2 = 8.29$. Notice also that this is the same data used to fit the bivariate normal distribution for which confidence ellipses are shown in Fig. 9.2.

The orientations of the two eigenvectors are shown in Fig. 9.3, although their lengths have been exaggerated for clarity. It is evident that the first eigenvector is aligned parallel to the direction in which the data jointly exhibit maximum variation. That is, the point cloud is inclined at the same angle as is $\mathbf{e}_1$. This angle is 32° from the horizontal (i.e., from the vector [1, 0]), according to Eq. (9.7). Since the data in this simple example exist in only $K = 2$ dimensions, the constraint that the second eigenvector must be perpendicular to the first determines its direction

up to sign (i.e., it could as easily be $-\mathbf{e}_2^T = [0.530, -0.848]$). This last eigenvector locates the direction in which data jointly exhibit their smallest variations.

The two eigenvectors determine an alternative coordinate system in which to view the data. This fact may become more clear if you rotate this book 32° to the right. Within this rotated coordinate system, each point is defined by a principal component vector $\mathbf{u}^T = [u_1, u_2]$ of new transformed variables, whose elements consist of the projections of the data onto the eigenvectors, according to the dot product in Eq. (9.36). Figure 9.3 illustrates this projection for the 15 January data point $\mathbf{x}'^T = [16.0, 17.8]$, which is indicated by the large square symbol. For this datum, $u_1 = (0.848)(16.0) + (0.530)(17.8) = 23.0$, and $u_2 = (-0.530)(16.0) + (0.848)(17.8) = 6.6$.

The sample variance of the new variable u_1 is an expression of the degree to which it spreads out along its axis (i.e., along the direction of $\mathbf{e}_1$). This dispersion is evidently greater than the dispersion of the data along either of the original axes, and in fact the sample variance of u_1 is equal to the eigenvalue $\lambda_1 = 254.76°\,\mathrm{F}^2$. While the points in the data set tend to exhibit quite different values of u_1, they have more similar values for u_2. That is, they are much less variable in the $\mathbf{e}_2$ direction, and the sample variance of u_2 is only $\lambda_2 = 8.29°\,\mathrm{F}^2$.

Since $\lambda_1 + \lambda_2 = s_{1,1} + s_{2,2} = 263.05°\,\mathrm{F}^2$, the new variables retain all of the variation exhibited by the original data. However, the fact that the point cloud seems to exhibit no slope in the new coordinate frame defined by the eigenvectors indicates that u_1 and u_1 are uncorrelated. Their lack of correlation can be verified by transforming the 31 pairs of minimum temperatures in Table A.1 to principal components and computing the Pearson correlation, which is zero. The variance–covariance matrix for the principal components is therefore $[\mathbf{\Lambda}]$, shown in Eq. (9.33).

The two original temperature variables are so strongly correlated that a very large fraction of their joint variance, $\lambda_1/(\lambda_1 + \lambda_2) = 0.968$, is represented by the first principal component. It would be said that the first principal component describes 96.8% of the total variance. The first principal component might be interpreted as reflecting the regional minimum temperature for the area including these two locations (they are about 50 miles apart), with the second principal component describing random variations departing from the overall regional value. ¤

9.3.2 Truncation of the Principal Components

Mathematically, there are as many eigenvectors of $[\mathbf{S}]$ as there are elements of the data vector $\mathbf{x}'$. That is, $M = K$. However, it is typical of atmospheric data that substantial covariances (or correlations) exist among the original K variables, and as a result there are few or no off-diagonal elements of $[\mathbf{S}]$ (or $[\mathbf{R}]$) that are near zero. This situation implies that there is redundant information in $\mathbf{x}$, and that the first few eigenvectors of its dispersion matrix will locate directions in which the

joint variability of the data is greater than the variability of any single element, of $\mathbf{x}'$. Similarly, the last few eigenvectors will point to directions in the K-dimensional space of $\mathbf{x}'$ where the data jointly exhibit very little variation. This feature was illustrated in Example 9.5 for daily temperature values measured at nearby locations.

To the extent that there is redundancy in the original data $\mathbf{x}'$, it is possible to capture most of their variations by considering only the most important directions of their joint variations. That is, most of the information content of the data may be represented using some smaller number $M^* < M$ of the principal components u_m. In effect, the original data set containing the K variables x_k is approximated by the smaller set of new variables u_m. If $M^* << M$, retaining only the first M^* of the principal components results in a much smaller data set. This "data compression" capability of PCA is often a primary motivation for its use.

The truncated representation of the original data can be expressed mathematically by the analysis formula, Eq. (9.37), but in this case the dimension of the truncated $\mathbf{u}$ is $(M^* \times 1)$, and $[\mathbf{E}]$ is the (nonsquare, $K \times M^*$) matrix whose columns consist only of the first M^* eigenvectors of $[\mathbf{S}]$. The synthesis formula, Eq. (9.38), is then only approximately true because the original data cannot be exactly resynthesized without using all M eigenvectors:

$$x'_k \approx \sum_{m=1}^{M^*} e_{km} u_m, \qquad k = 1, \ldots, K. \tag{9.39}$$

Regardless of the choice of how many principal components are retained, the proportion of the total joint variation in the data represented by the M^* retained principal components can be computed from the eigenvalues as

$$\text{Proportion of total variance described} = \frac{\sum_{m=1}^{M^*} \lambda_m}{\sum_{m=1}^{M} \lambda_m} = \frac{\sum_{m=1}^{M^*} \lambda_m}{\sum_{k=1}^{K} s_{k,k}}, \tag{9.40}$$

where $s_{k,k}$ is the kth diagonal element of $[\mathbf{S}]$, or the sample variance of x'_k.

An important application of the truncation of the full set of M principal components is in the setting of multiple regression (Chapter 6). It often occurs in multiple regression that the predictor (x) variables exhibit very large mutual correlations, which is referred to as *multicollinearity*. In this situation, it turns out that the sampling distributions of the estimated regression coefficients can become very broad, with the practical consequence that a forecasting equation may perform very badly when implemented on future data independent of the training sample. The problem can be rectified by first subjecting the predictor variables to a PCA, and then using the first M^* principal components $\mathbf{u}$ as predictors in place of the original variables $\mathbf{x}$. This approach is called *principal-components regression*. Because the principal components are uncorrelated, they do not exhibit the multicollinearity problem, and their regression coefficients will be estimated with

much better precision. As is the case with predictors that are harmonic functions [Eq. (8.69)], principal-component predictors can be added or deleted from a candidate regression without changing the regression coefficients of other principal-component predictors. This property can aid in the understanding of regression results if the principal components themselves admit of physical interpretations. If M^* is chosen to be large enough, most of the information contained jointly in the full set of the original correlated variables will also be represented in the principal components, so that, in general, little information will be sacrificed. Principal components regression is discussed more fully in Jolliffe (1986) and Draper and Smith (1981).

9.3.3 How Many Principal Components Should Be Retained?

The question, of course, arises as to how few principal components can be retained without discarding important information carried in the original data. There is not a single clear criterion that can be used to choose this number M^* of principal components that are best retained in a given circumstance. While the choice of the truncation level can be aided by one or more of the many available principal-component selection rules, it is ultimately a subjective choice that will depend in part on the data at hand and the purposes of the PCA.

Many of the established principal component selection rules are rooted in the statistical literature, but a substantial amount of work has been done in this area with a specific orientation toward atmospheric and oceanographic data sets (especially Preisendorfer *et al.*, 1981). In this latter context the problem is often viewed in the electronic signal processing paradigm of separating "signal" from "noise." That is, one imagines that there are some number of distinct signals, or modes of variation, embedded in the data, and that this number (M^*) is less than the number of measurements (K) composing each data vector $\mathbf{x}$. In this view, the goal is then to find and discard the $K - M^*$ "noise" components, which are regarded as random-number contaminants that have been added to the interesting data (the "signal"). It is hoped that selection of the first M^* principal components will "filter" the noise and leave behind the meaningful portion of the original data.

Most principal-component selection rules can be viewed as *dominant variance* rules, in that it is the M^* principal components with the largest eigenvalues (which therefore represent the largest fraction of variance in the original data) that are selected. The differences among these rules are with respect to how and where the line is drawn between the retained and discarded principal components.

Figure 9.4 illustrates two graphically based principal-component selection rules for a PCA involving $K = 6$ observations in each data vector $\mathbf{x}$. Figure 9.4a shows the plot of the eigenvalues vs. the corresponding principal component number, called the *scree graph* (or, sometimes, the "scree test"). Figure 9.4b shows the same data with the eigenvalues plotted logarithmically, which is called the *log-eigenvalue* (LEV) *diagram*. Use of either the scree graph or the LEV diagram

Fig. 9.4 Graphical displays of eigenvalue magnitudes as a function of the principal-component number (heavier lines connecting circled points), for a $\mathbf{K} = 6$-dimensional analysis: (a) linear scaling, or "scree" graph; (b) logarithmic scaling, or LEV diagram. Both the scree and LEV criteria would lead to retention of the first three principal components in this analysis. Lighter lines in both panels show results of the resampling tests necessary to apply rule N of Preisendorfer *et al.* (1981). Dashed line is median of eigenvalues for 1000 (6×6) dispersion matrices of independent Gaussian variables, constructed using the same sample size as the data being analyzed. Solid lines indicate the 5th and 95th percentiles of these simulated eigenvalue distributions. Rule N would indicate retention of only the first two principal components, as these are the ones significantly larger than what would be expected from data with no correlation structure.

requires a subjective judgment about the existence and location of a "break" in the plotted curve.

When using the scree graph qualitatively, the goal is to locate a point separating a steeply sloping portion to the left, and a more shallowly sloping portion to the right. The principal-component number at which the separation occurs is then taken as M^*. In Figure 9.4a this point would probably be chosen as $M^* = 3$. On the other hand, the LEV diagram is motivated by the idea that, if the last $K - M^*$ principal components represent uncorrelated noise, then the magnitudes of their eigenvalues should decay exponentially with increasing principal-component number. This behavior should be identifiable in the logarithmic LEV diagram as a straight-line portion on its right-hand side. The M^* retained principal components would then be the ones whose log-eigenvalues lie above the leftward extrapolation of this line. In Fig. 9.4b, $M^* = 3$ would probably be chosen by most viewers of this LEV diagram, although this choice is not unambiguous. Similarly, not all scree graphs exhibit an obvious "elbow."

Another group of dominant-variance selection rules involving comparing each eigenvalue (and therefore the variance described by its principal component) to the amount of the joint variance reflected in the "average" eigenvalue. Principal

components whose eigenvalues are above this threshold are retained. This set of selection rules can be summarized by the criterion

$$\text{Retain } \lambda_m \quad \text{if} \quad \lambda_m > \frac{T}{K} \sum_{k=1}^{K} s_{k,k}, \tag{9.41}$$

where $s_{k,k}$ is the sample variance of the kth element of $\mathbf{x}$. The simplest of these, called *Kaiser's rule*, retains the principal components accounting for more than the average amount of the total variance. Thus, Kaiser's rule uses Eq. (9.41) with the threshold parameter $T = 1$. Jolliffe (1972) has argued that Kaiser's rule is too strict, and suggested the alternative of $T = 0.7$. A third alternative is to use the *broken stick model*, so called because it is based on the expected length of the mth longest piece of a randomly broken unit line segment. According to this criterion, the threshold parameter in Eq. (9.41) is taken to be

$$T = \frac{1}{M} \sum_{j=m}^{M} \frac{1}{j}. \tag{9.42}$$

This rule yields a different threshold for each principal component: $T = T(m)$. All three criteria described above lead to choosing $M^* = 2$ for the data in Fig. 9.4.

Of the many principal-component selection rules devised and investigated by Preisendorfer *et al.* (1981), the most commonly used is their *rule N*. Rule N identifies the largest M^* principal components to be retained on the basis of resampling (or Monte Carlo) tests, involving the distribution of eigenvalues of randomly generated dispersion matrices. The procedure involves repeatedly generating sets of vectors of independent Gaussian random numbers with the same dimension (K) and sample size (n) as the data $\mathbf{x}$ being analyzed, and then computing the eigenvalues of their dispersion matrices. These randomly generated eigenvalues are then scaled in a way that makes them comparable to the eigenvalues λ_m to be tested, for example, by requiring that the sum of each set of randomly generated eigenvalues will equal the sum of the eigenvalues computed from the original data. Each λ_m from the real data is then compared to the empirical distribution of its synthetic counterparts, and is retained if it is larger than 95% of these.

The light lines in the panels of Fig. 9.4 illustrate the use of rule N to select a principal-component truncation level. The dashed lines reflect the medians of 1000 sets of eigenvalues computed from 1000 (6×6) dispersion matrices of independent Gaussian variables, constructed using the same sample size as the data being analyzed. The solid lines show 95th and 5th percentiles of those distributions for each of the six eigenvalues. The first two eigenvalues λ_1 and λ_2 are larger than more than 95% of their synthetic counterparts, and for these the null hypothesis that the corresponding principal components represent only "noise" would therefore be rejected at the 5% level. Accordingly, rule N would choose $M^* = 2$ for this data. As an aside, note that the nearly straight-line relationships exhibited by the synthetic eigenvalue quantiles in Fig. 9.4b support the premise of

the LEV criterion, that the eigenvalues of dispersion matrices of uncorrelated noise should decrease exponentially.

A table of 95% critical values for rule N, for selected sample sizes n and dimensions K, is presented in Overland and Preisendorfer (1982). Corresponding large-sample tables are given in Preisendorfer *et al.* (1981) and Preisendorfer (1988). Preisendorfer (1988) notes that if there is substantial autocorrelation (i.e., time correlation) present in the individual variables x_k, that it may be more appropriate to construct the resampling distributions for rule N (or to use the tables just mentioned) using the smallest effective sample size [Eq. (5.12)] among the x_k, rather than using n independent vectors of Gaussian variables to construct each synthetic dispersion matrix. Another potential problem with rule N, and other similar procedures, is that the data $\mathbf{x}$ may not be approximately Gaussian. For example, one or more of the x_k values could be precipitation variables. To the extent that the original data are not Gaussian, the resampling procedure will not simulate accurately the physical process that generated them, and the results of the tests may be misleading. A possible remedy for the problem of non-Gaussian data might be to use a bootstrap version of rule N, although this approach seems not to have been tried in the literature to date.

9.3.4 PCA Based on the Covariance Matrix versus the Correlation Matrix

PCA can be conducted as easily on the correlation matrix $[\mathbf{R}]$ as it can on the covariance matrix $[\mathbf{S}]$. The correlation matrix is the variance–covariance matrix of the vector of standardized variables $\mathbf{z}$ [Eq. (9.21)]. The vector of standardized variables $\mathbf{z}$ is related to the vectors of original variables $\mathbf{x}$ and centered variables $\mathbf{x}'$ according to

$$\mathbf{z} = [\mathbf{D}]^{-1}\mathbf{x}' = [\mathbf{D}]^{-1}(\mathbf{x} - \bar{\mathbf{x}})$$

$$= \begin{bmatrix} \frac{1}{\sqrt{s_{1,1}}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{2,2}}} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{s_{3,3}}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{s_{K,K}}} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ x_3 - \bar{x}_3 \\ \vdots \\ x_K - \bar{x}_K \end{bmatrix}, \tag{9.43}$$

where the diagonal matrix $[\mathbf{D}]$ was defined in conjunction with Eq. (9.20), and contains diagonal elements that are the sample standard deviations of x_k (and of x'_k). Therefore, PCA on the correlation matrix amounts to analysis of the joint variance structure of the standardized variables z_k, as computed using either Eq. (9.43) or (in scalar form) Eq. (3.16).

The difference between PCAs performed using the variance–covariance and correlation matrices will be one of emphasis. Since the PCA seeks to find variables successively maximizing proportion of the total variance, $\Sigma_k s_{k,k}$, represented, analyzing the covariance matrix **[S]** results in principal components that emphasize the x_k values having the largest variances. Other things equal, the tendency will be for the first few eigenvectors to align near the directions of the variables having the biggest variances. In Example 9.5, the first eigenvector points more toward the Ithaca minimum temperature axis because the variance of the Ithaca minimum temperatures is larger than the variance of the Canandaigua minimum temperatures. Conversely, PCA applied to the correlation matrix **[R]** weights all of the standardized variables z_k equally, since all have equal (unit) variance.

If the PCA is conducted using the correlation matrix, the analysis formula, Eqs. (9.36) and (9.37), will pertain to the standardized variables, z_k and **z**, respectively. Similarly, the synthesis formula, Eq. (9.38), will pertain to **z** and z_k rather than to $\mathbf{x}'$ and x'_k. In this case the original data **x** can be recovered from the result of the synthesis formula by reversing the standardization given by Eq. (9.43):

$$\mathbf{x} = [\mathbf{D}]\mathbf{z} + \bar{\mathbf{x}}. \tag{9.44}$$

Although **z** and $\mathbf{x}'$ can be easily obtained from each other using Eq. (9.43), the eigenvalue/eigenvector pairs of **[R]** and **[S]** do not bear simple relationships to one another. In general, it is not possible to compute the principal components of one knowing only the principal components of the other. This fact implies that these two alternatives for PCA do not yield equivalent information, and that it is important to make an intelligent choice of one over the other for a given application. If an important goal of the analysis is to identify or isolate the strongest variations in a data set, the better choice will usually be PCA using the covariance matrix. This will be particularly so if all the x_k values are measured in the same physical units. In this situation, the choice between PCA of the covariance or the correlation matrix will depend on the judgment of the analyst and the purpose of the study. For example, in analyzing gridded numbers of extratropical cyclones, Overland and Preisendorfer (1982) found that PCA on their covariance matrix better identified regions having the highest variability in cyclone numbers, whereas correlation-based PCA was more effective at locating the primary storm tracks.

However, if the analysis is of unlike variables, that is, variables not measured in the same units, it will almost always be preferable to compute the PCA using the correlation matrix. Measurement in unlike physical units yields arbitrary relative scalings of the variables, which results in arbitrary relative magnitudes of the variances of these variables. To take a simple example, the variance of a set of temperatures measured in degrees Fahrenheit will be $(1.8)^2 = 3.24$ times as large as the variance of the same temperatures expressed in degrees Celsius. If the PCA has been done using the correlation matrix, the analysis formula, Eq. (9.37), pertains to the vector **z** rather than $\mathbf{x}'$; and the truncated synthesis in Eq. (9.39) will

yield approximately the standardized variables z_k. The summations in the denominators of Eq. (9.40) will equal the number of standardized variables since each has unit variance.

Example 9.6. Correlation-versus-Covariance-Based PCA for Arbitrarily Scaled Quantities

The importance of basing a PCA on the correlation matrix when the variables being analyzed are not measured on comparable scales is illustrated in Table 9.2. This table summarizes PCAs of the January 1987 data in Table A.1 in (a) unstandardized (covariance matrix) and (b) standardized (correlation matrix) forms. Sample variances of the variables are shown, as are the six eigenvectors, the six eigenvalues, and the cumulative percentages of variance accounted for by the principal components. The (6 × 6) arrays in the upper-right portions ("Canan-

Table 9.2

Comparison of PCA Computed Using (a) the Covariance Matrix and (b) the Correlation Matrix of the Data in Table A.1[a]

Variable	Sample variance	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$	$\mathbf{e}_6$
		a. Covariance Results					
Ithaca ppt	0.059 in.^{2}	.003	.017	.002	−.028	.818	−.575
Ithaca T_{max}	892.2°F^2	.359	−.628	.182	−.665	−.014	−.003
Ithaca T_{min}	185.5°F^2	.717	.527	.456	.015	−.014	.000
Canandaigua ppt	0.028 in^2	.002	.010	.005	−.023	.574	.818
Canandaigua T_{max}	61.8°F^2	.381	−.557	.020	.737	.037	.000
Canandaigua T_{min}	77.6°F^2	.459	.131	−.871	−.115	−.004	.003
Eigenvalues, λ_k		337.7	36.9	7.49	2.38	0.065	0.001
Cumulative % variance		87.8	97.4	99.3	99.9	100.0	100.0
		b. Correlation Results					
Ithaca ppt	1.000	.142	.677	.063	−.149	−.219	.668
Ithaca T_{max}	1.000	.475	−.203	.557	.093	.587	.265
Ithaca T_{min}	1.000	.495	.041	−.526	.688	−.020	.050
Canandaigua ppt	1.000	.144	.670	.245	.096	.164	−.658
Canandaigua T_{max}	1.000	.486	−.220	.374	−.060	−.737	−.171
Canandaigua T_{min}	1.000	.502	−.021	−.458	−.695	−.192	−.135
Eigenvalues, λ_k		3.532	1.985	0.344	0.074	0.038	0.027
Cumulative % variance		58.9	92.0	97.7	98.9	99.5	100.0

[a] The sample variances of each variable are shown, as are the six eigenvectors $\mathbf{e}_m$ arranged in decreasing order of their eigenvalues λ_m. The cumulative percentage of variance represented is calculated according to Eq. (9.40) multiplied by 100%. The much smaller variances of the precipitation (ppt) variables are artifacts of the measurement units, but result in precipitation being unimportant in the first four principal components computed from the covariance matrix, which collectively account for 99.9% of the total variance of the data set. Computing the principal components from the correlation matrix ensures that variations of the temperature and precipitation variables are weighted equally.

daigua" and "Ithaca" rows, columns 3–8) of this table constitute the matrix of eigenvectors, **[E]**, in each case.

Because of the magnitudes of the variations of the data in relation to their measurement units, the variances of the unstandardized precipitation data are tiny in comparison to the variances of the temperature variables. This is purely an artifact of the measurement unit for precipitation (inches) being relatively large in comparison to the range of variation of the data (about 1 in.), and the measurement unit for temperature (°F) being relatively small in comparison to the range of variation of the data (about 40°F). If the measurement units had been millimeters and degrees Celsius, respectively, the differences in variances would have been much smaller. If the precipitation had been measured in micrometers, the variances of the precipitation variables would dominate the variances of the temperature variables.

Because the variances of the temperature variables are so much larger than the variances of the precipitation variables, the PCA of the covariance matrix is dominated by the temperatures. The eigenvector elements corresponding to the two precipitation variables are negligibly small in the first four eigenvectors, so these variables make negligible contributions to the first four principal components. However, these first four principal components collectively describe 99.9% of the joint variance. No reasonable principal-component truncation rule would retain the last two principal components, which carry almost all of the precipitation information. An application of the truncated synthesis formula, Eq. (9.39), would therefore result in reconstructed precipitation values very near their average values. That is, essentially none of the variation in precipitation would be represented.

Since the correlation matrix is the variance–covariance matrix for comparably scaled variables z_k, each has equal variance. Unlike the analysis on the covariance matrix, the PCA does not ignore the precipitation variables when the correlation matrix is analyzed. Here the first (and most important) principal component represents primarily the closely intercorrelated temperature variables, as can be seen from the relatively larger elements of $\mathbf{e}_1$ for the temperatures. However, the second principal component, which accounts for 33.1% of the total variance in the scaled data set, represents primarily the precipitation variations. The eigenvalues for this PCA are shown in Fig. 9.4. The first two principal components were retained according to all of the selection rules discussed in the previous section, and would almost surely be retained in any reasonable truncation rule. The result would be that the precipitation variations would not be lost in the truncated data representation, but rather would be very nearly completely reconstructed by an application of Eq. (9.39). ¤

9.3.5 Application of PCA to Fields

The overwhelming majority of applications of PCA to atmospheric data have involved analysis of fields (i.e. spatial arrays of variables) such as geopotential

heights, temperatures, and precipitation. In these cases the full data set consists of multiple observations of a field or of a set of fields. Frequently these multiple observations take the form of time series, for example, a sequence of daily hemispheric 500-mb heights. Another way to look at this kind of data is as a collection of K mutually correlated time series that have been sampled at each of K gridpoints or station locations. The goal of PCA as applied to this type of data is usually to explore, or to express succinctly, the joint space/time variations of the many variables in the data set.

Even though the locations at which the field is sampled are spread over a two-dimensional (or possibly three-dimensional) space, the data from these locations at a given observation time are arranged in the one-dimensional vector $\mathbf{x}$. That is, regardless of their geographic arrangement, each location is assigned a number (as in Fig. 7.11) from 1 to K, which refers to the appropriate element in the data vector $\mathbf{x}^{\mathrm{T}} = [x_1, x_2, x_3, \ldots, x_K]$. In this most common application of PCA to fields, the data matrices $[\mathbf{X}]$ and $[\mathbf{X}']$ are thus dimensioned $(n \times K)$, or (time $\times$ space), since data at K locations in space have been sampled at n different times.

To emphasize that the original data consists of K time series, the analysis equation [(9.36) or (9.37)] is sometimes written with an explicit time index:

$$\mathbf{u}(t) = [\mathbf{E}]^{\mathrm{T}}\mathbf{x}'(t), \tag{9.45a}$$

or, in scalar form,

$$u_m(t) = \sum_{k=1}^{K} e_{km} x'_k(t), \qquad m = 1, \ldots, M. \tag{9.45b}$$

Here the time index t runs from 1 to n. The synthesis equation [(9.38) or (9.39)] can be written using the same notation. Equation (9.45) emphasizes that, if the data $\mathbf{x}$ consist of a set of time series, then the principal components $\mathbf{u}$ are also time series. The time series of one of the principal components, $u_m(t)$, may very well exhibit serial correlation (correlation with itself through time), and the principal-component time series are sometimes analyzed using tools presented in Chapter 8. However, each principal-component time series will be uncorrelated with the time series of all the other principal components.

When the K elements of $\mathbf{x}$ are measurements at different locations in space, the eigenvectors can be displayed graphically in a quite informative way. Notice that each eigenvector contains exactly K elements, and that these elements have a one-to-one correspondence with each of the K locations in the dot product from which the corresponding principal component is calculated [Eq. (9.45b)]. Each eigenvector element e_{km} can be plotted on a map at the same location as its corresponding data value x'_k, and this "field" of eigenvector elements can itself be displayed with smooth contours in the same way as ordinary meteorological fields. Such maps depict clearly which locations are contributing most strongly to the respective principal components. Looked at another way, such maps indicate the geographic distribution of simultaneous data anomalies represented by the corre-

sponding principal component. These geographic displays of eigenvectors are also sometimes interpreted as representing uncorrelated modes of variability of the field from which the PCA was extracted.

Figure 9.5, from Wallace and Gutzler (1981), shows the first four eigenvectors of a PCA of the correlation matrix for winter monthly-mean 500-mb heights at gridpoints in the Northern Hemisphere. The numbers below and to the right of the panels show the percentage of the total hemispheric variance [Eq. (9.40) × 100%] represented by each of the corresponding principal components. Together, the first four principal components account for nearly half of the (normalized) hemispheric winter height variance. These patterns resemble the *teleconnectivity* patterns for the same data shown in Fig. 3.20, and apparently reflect the same underlying physical processes in the atmosphere. For example, Fig. 9.5b evidently reflects the "PNA" pattern of alternating height anomalies stretching from the Pacific Ocean to northwestern North America to southeastern North America. A positive value of the second principal component of this data set corresponds to negative 500-mb height anomalies (troughs) in the northwestern Pacific and in the southeastern

Fig. 9.5 Spatial displays of the first four eigenvectors of gridded winter monthly-mean 500-mb heights for the Northern Hemisphere, 1962–1977. This PCA was computed using the correlation matrix of the height data. Percentage values below and to the right of each map are proportion of total variance [Eq. (9.40)] × 100%. The patterns resemble the "teleconnectivity" patterns for the same data (Fig. 3.20). [From Wallace and Gutzler (1981), reproduced with permission of the American Meteorological Society.]

Fig. 9.5 *Continued.*

Fig. 9.5 *Continued.*

United States, and to positive height anomalies (ridges) the western part of the continent, and over the central tropical Pacific. A negative value of the second principal component yields the reverse pattern of anomalies, and a more zonal 500-mb flow over North America.

It is also possible to apply PCA to vector-valued fields, which are fields with observations of more than one variable at each location or gridpoint. This is equivalent to simultaneous analysis of two or more fields. If there are L such variables at each of the K gridpoints, then the dimensionality of the data vector $\mathbf{x}$ is given by the product KL. The first K elements of $\mathbf{x}$ are observations of the first variable, the second K elements are observations of the second variable, and the last K elements of $\mathbf{x}$ will be observations of the Lth variable. Since the L different variables will generally be measured in unlike units, it will almost always be desirable to base the PCA of such data on the correlation matrix. The dimension of $[\mathbf{R}]$, and of the matrix of eigenvectors $[\mathbf{E}]$, will then be $(M \times M)$, where $M = KL$. Application of PCA to this kind of correlation matrix will produce principal components successively maximizing the joint variance of the L variables in a way that considers the correlations both between and among these variables at the K locations.

Figure 9.6 illustrates the structure of the correlation matrix (left) and the matrix of eigenvectors (right) for PCA of vector field data. The first K rows of $[\mathbf{R}]$ contain

Fig. 9.6 Structures of the correlation matrix and the matrix of eigenvectors for PCA of vector field data. The basic data consist of multiple observations of L variables at each of K locations, so the dimensions of both **[R]** and **[E]** are $(M \times M)$, where $M = KL$. The correlation matrix consists of $(K \times K)$ submatrices containing the correlations between sets of the L variables jointly at the K locations. The submatrices located on the diagonal of **[R]** thus contain ordinary correlation matrices for each of the L variables. The off-diagonal submatrices contain correlation coefficients, but are not symmetrical and will not contain 1's on their diagonals. Each eigenvector column of **[E]** similarly consists of L segments, each of which contains K elements pertaining to the individual locations.

the correlations between the first of the L variables at these locations and all of the M variables. Rows $K + 1$ to $2K$ similarly contain the correlations between the second of the L variables and all the M variables, and so on. Another way to look at the correlation matrix is as a collection of L^2 submatrices, each dimensioned $(K \times K)$, which contain the correlations between sets of the L variables jointly at the K locations. The submatrices located on the diagonal of **[R]** thus contain ordinary correlation matrices for each of the L variables. The off-diagonal submatrices contain correlation coefficients, but are not symmetric and will not contain 1's on their diagonals. However, the overall symmetry of **[R]** implies that $[\mathbf{R}_{i,j}] = [\mathbf{R}_{j,i}]^{\mathrm{T}}$. Similarly, each column of **[E]** consists of L segments, and each of these segments contains the **K** elements pertaining to each of the individual locations.

The eigenvector elements resulting from a PCA of a vector field can be displayed graphically in a manner that is similar to the maps drawn for ordinary, scalar fields. Here, each of the L groups of K eigenvector elements is either overlaid on the same base map, or plotted on separate maps. Figure 9.7, from Kutzbach (1967), illustrates this process for the case of $L = 2$ observations at each location. The two variables are average January surface pressure and average January temperature, measured at $K = 23$ locations in North America. The heavy lines are an analysis of the (first 23) elements of the first eigenvector that pertain to the pressure data, and the dashed lines with shading show an analysis of the temperature (second 23) elements of the same eigenvector. The corresponding principal component accounts for 28.6% of the total variance of the $23 \times 2 = 46$ variables.

In addition to effectively condensing an enormous amount of information, the patterns shown in Fig. 9.7 are physically consistent with atmospheric processes. In particular, the temperature anomalies are consistent with patterns of thermal

Fig. 9.7 Spatial display of the elements of the first eigenvector of the (46 × 46) correlation matrix of average January sea-level pressures and temperatures at 23 locations in North America. The first principal component of this correlation matrix accounts for 28.6% of the joint (scaled) variance of the pressures and temperatures. Heavy lines are a hand analysis of the sea-level pressure elements of the first eigenvector, and dashed lines with shading are a hand analysis of the temperature elements of the same eigenvector. The joint variations of pressure and temperature depicted are physically consistent with temperature advection in response to the pressure anomalies. [From Kutzbach (1967), reproduced with permission of the American Meteorological Society.]

advection implied by the pressure anomalies. If the first principal component u_1 is positive for a particular month, the solid contours imply positive pressure anomalies in the north and east, with lower than average pressures in the southwest. On the west coast, this pressure pattern would result in weaker than average westerly surface winds and stronger than average northerly surface winds. The

resulting advection of cold air from the north would produce colder temperatures, and this cold advection is reflected by the negative temperature anomalies in this region. Similarly, the pattern of pressure anomalies in the southeast would enhance southerly flow of warm air from the Gulf of Mexico, resulting in positive temperature anomalies as shown. Conversely, if u_1 is negative, reversing the signs of the pressure eigenvector elements implies enhanced westerlies in the west, and northerly wind anomalies in the southeast, which are consistent with positive and negative temperature anomalies, respectively. These temperature anomalies are indicated by Fig. 9.7, when the signs on the temperature contours are also reversed.

Figure 9.7 is a simple example involving familiar variables. Its interpretation is easy and obvious if one is conversant with the climatological relationships of pressure and temperature patterns over North America in winter. However, the physical consistency exhibited in this example (where the "right" answer is known ahead of time) is indicative of the power of this kind of PCA to uncover meaningful joint relationships among atmospheric (and other) fields in an exploratory setting, where clues to possibly unknown underlying physical mechanisms may be hidden in the complex relationships among several fields.

9.3.6 Rotation of the Eigenvectors

When the PCA eigenvector elements are plotted geographically, there is a strong tendency to try to ascribe physical interpretations to the corresponding principal components. The results shown in Fig. 9.7 indicate that it can be both appropriate and informative to do so. However, the orthogonality constraint on the eigenvectors [Eq. (9.27)] can lead to problems with these interpretations, especially for the second and subsequent principal components. While the orientation of the first eigenvector in its M-dimensional space is determined solely by the direction of the maximum variation in the data, subsequent vectors must be orthogonal to previously determined eigenvectors, regardless of the nature of the physical processes that gave rise to the data. To the extent that the underlying physical processes are not independent, interpretation of the corresponding principal components as being independent "modes of variability" will not be justified. Although the first principal component may represent an important mode of variability or physical process, it may well also include aspects of other correlated modes or processes. Thus, the orthogonality constraint on the eigenvectors can result in the influences of several distinct physical processes being jumbled together in a single principal component.

When physical interpretation, rather than data compression, is a primary goal of PCA, it is often desirable to rotate a subset of the initial eigenvectors to a second set of new coordinate vectors. Usually it is some number M^* of the leading eigenvectors (i.e. eigenvectors with largest corresponding eigenvalues) of the original PCA that are rotated, with M^* determined using a truncation criterion such as

Eq. (9.41). The review article by Richman (1986) explains in some detail why rotation may be useful, and also gives a technical exposition of the subject.

As a consequence of rotation of the eigenvectors, a second set of new variables is produced, called *rotated principal components*. The rotated principal components are obtained from the original data in a manner similar to the PCA analysis equation, as the dot product of data vectors and the rotated eigenvectors. Depending on the method used to rotate the eigenvectors, the resulting rotated principal components may or may not be mutually uncorrelated.

A number of procedures for rotating the original eigenvectors exist, but all seek to produce what is known as *simple structure* in the resulting analysis. Roughly speaking, simple structure is achieved if a large number of the elements of the resulting rotated vectors are near zero, and few of the remaining elements correspond to (have the same index k as) elements that are also not near zero in the other rotated vectors. The result is that the rotated vectors represent mainly the few original variables corresponding the elements not near zero, and that the representation of the original variables is split between as few of the rotated principal components as possible.

Figure 9.8, from Horel (1981), shows spatial displays of the first two rotated eigenvectors of monthly-averaged hemispheric winter 500-mb heights. Using the truncation criterion of Eq. (9.41) with $T = 1$, the first 19 eigenvectors of the correlation matrix of these data were rotated. The two patterns in Fig. 9.8 are similar to the first two unrotated eigenvectors derived from the same data (Figs. 9.5a and 9.5b), although the signs have been (arbitrarily) reversed. However, the rotated vectors conform more to the idea of simple structure in that more of the hemispheric fields are fairly flat (near zero) in Fig. 9.8, and each panel emphasizes more uniquely a particular feature of the variability of the 500-mb heights corresponding to the teleconnection patterns in Fig. 3.20. The rotated vector in Fig. 9.8a focuses primarily on height differences in the northwestern and western tropical Pacific, called the *western Pacific* teleconnection pattern. It thus represents variations in the 500-mb jet at these longitudes, with positive values of the corresponding rotated principal component indicating weaker-than-average westerlies, and negative values indicating the reverse. Similarly, the PNA pattern stands out exceptionally clearly in Fig. 9.8b, where the rotation has separated it from the eastern hemisphere pattern evident in Fig. 9.5b.

9.3.7 The Varied Terminology of PCA

The subject of PCA is sometimes regarded as a difficult and confusing one, but much of this confusion derives from a proliferation of the associated terminology, especially in writings by analysts of atmospheric data. Table 9.3 organizes the more common of these in a way that may be helpful in deciphering the PCA literature.

Lorenz (1956) introduced the term *empirical orthogonal function* (EOF; men-

Fig. 9.8 Spatial displays of the first two rotated eigenvectors of monthly-averaged hemispheric winter 500-mb heights. The data is the same as that underlying Fig. 9.6, but the rotation has better isolated the patterns of variability, allowing a clearer interpretation in terms of the teleconnection patterns in Fig. 3.20. [From Horel (1981), reproduced with permission of the American Meteorological Society.]

Table 9.3

A Partial Guide to Synonymous Terminology Associated with PCA

Eigenvectors, $\mathbf{e}_m$	Eigenvector elements, $e_{k,m}$	Principal components, $\mathbf{u}_m$	Principal-component elements, $u_{i,m}$
EOFs	Loadings	Empirical orthogonal variables	Scores
Modes of variation	Coefficients		Amplitudes
Pattern vectors	Pattern coefficients		Expansion coefficients
Principal axes	Empirical orthogonal weights		Coefficients
Principal vectors			
Proper functions			
Principal directions			

tioned earlier) into the literature as another name for the eigenvectors of a PCA. The terms *modes of variation* and *pattern vectors* are also used primarily by analysts of geophysical data, especially in relation to analysis of fields as described previously. The remaining terms for the eigenvectors derive from the geometric interpretation of the eigenvectors as basis vectors, or axes, in the K-dimensional space of the data. These terms are used in the literature of a broader range of disciplines.

The most common name for individual elements of the eigenvectors in the statistical literature is *loading*, connoting the weight of the kth variable x_k that is borne by the mth eigenvector $\mathbf{e}_m$ through the individual element e_{km}. The term *coefficient* is also a usual one in the statistical literature. The term *pattern coefficient* is used mainly in relation to PCA of field data, where the spatial patterns exhibited by the eigenvector elements can be illuminating. *Empirical orthogonal weights* is a term that is sometimes used to be consistent with the naming of the eigenvectors as EOFs.

The new variables u_m defined with respect to the eigenvectors are almost universally called *principal components*. However, they are sometimes known as *empirical orthogonal variables* when the eigenvectors are called EOFs. There is more variation in the terminology for the individual values of the principal components $u_{i,m}$ corresponding to particular data vectors $\mathbf{x}'_i$ [or, in the notation of Eq. (9.45), of $u_m(t)$ corresponding to a particular $\mathbf{x}'(t)$]. In the statistical literature these are most commonly called *scores*, which has a historical basis in the early and widespread use of PCA in psychometrics. In atmospheric applications, the principal component elements are often called *amplitudes* by analogy to the amplitudes of a Fourier series, that multiply the (theoretical orthogonal) sine and cosine functions. Similarly, the term *expansion coefficient* is also used for this meaning. Sometimes expansion coefficient is shortened simply to *coefficient*, al-

though this can be the source of some confusion since it is more standard for the term *coefficient* to denote an eigenvector element.

9.3.8 Scaling Conventions in PCA

Another contribution to confusion in the literature of PCA is the existence of alternative scaling conventions for the eigenvectors. The presentation in this section has assumed that the eigenvectors are scaled to unit length: $\|\mathbf{e}_m\| \equiv 1$. Recall that vectors of any length will satisfy Eq. (9.25) if they point in the appropriate direction, and as a consequence it is common for the output of eigenvector computations to be expressed with this scaling.

However, it is sometimes useful to express and manipulate PCA results using alternative scalings of the eigenvectors. When this is done, each element of an eigenvector is multiplied by the same constant, so their relative magnitudes and relationships remain unchanged. Thus, for example, the overall patterns in maps such as those shown in Figs. 9.5, 9.7, and 9.8 are unchanged under different scalings, and it is only the magnitudes of the contour labels that change. Therefore, the qualitative results of an exploratory analysis based on PCA do not depend on the scaling selected, but if different, related analyses are to be compared it is important to be aware of the scaling convention used in each.

Rescaling the lengths of the eigenvectors changes the magnitudes of the principal components correspondingly. That is, multiplying the eigenvector $\mathbf{e}_m$ by a constant requires that the principal component scores u_m be multiplied by the same constant in order for the analysis formulas that define the principal components [Eqs. (9.36) and (9.37)] to remain valid. The expected values of the principal component scores for centered data $\mathbf{x}'$ are zero, and multiplying the principal components by a constant will produce rescaled principal components whose means are also zero. However, their variances will change by a factor of the square of the scaling constant, and the magnitudes of their correlations with the original variables will be affected as well.

Table 9.4 summarizes the effects of three common scalings of the eigenvectors on the properties of the principal components. The first row indicates their properties under the scaling convention $\|\mathbf{e}_m\| \equiv 1$ adopted in this presentation. The eigenvectors plotted in Fig. 9.8 conform to this scaling. Under this scaling, the expected value (mean) of each of the principal components is zero, and the variance of each is equal to the respective eigenvalue, λ_m. This result is simply an expression of the diagonalization of the variance–covariance matrix [Eq. (9.29)] produced by adopting the geometric coordinate system defined by the eigenvectors. When scaled in this way, the correlation between a principal component u_m and a variable x_k is proportional to the eigenvector element that connects them, $e_{k,m}$, with the constant of proportionality being the ratio of the square root of the corresponding eigenvalue $(\lambda_m)^{1/2}$ to the standard deviation of the variable s_k. The

Table 9.4

Three Common Eigenvector Scalings Used in PCA and Their Consequences for the Properties of the Principal Components u_m, and Relationship to Original Variables x_k and Standardized Original Variables z_k

Eigenvector scaling	$E[u_m]$	$\mathrm{Var}[u_m]$	$\mathrm{Corr}[u_m, x_k]$	$\mathrm{Corr}[u_m, z_k]$
$\lVert \mathbf{e}_m \rVert = 1$	0	λ_m	$e_{k,m}(\lambda_m)^{1/2}/s_k$	$e_{k,m}(\lambda_m)^{1/2}$
$\lVert \mathbf{e}_m \rVert = (\lambda_m)^{1/2}$	0	λ_m^2	$e_{k,m}/s_k$	$e_{k,m}$
$\lVert \mathbf{e}_m \rVert = (\lambda_m)^{-1/2}$	0	1	$e_{k,m}\lambda_m/s_k$	$e_{k,m}\lambda_m$

correlation between u_m and the standardized variable z_k is given by the product of the eigenvector element and the square root of the eigenvalue, since the standard deviation of a standardized variable is one.

The eigenvectors are commonly rescaled by multiplying each element by the square root of the corresponding eigenvalue. This rescaling produces vectors of differing lengths, $\lVert \mathbf{e}_m \rVert \equiv (\lambda_m)^{1/2}$, but that point in exactly the same directions as the original eigenvectors with unit lengths. The eigenvectors plotted in Figs. 9.5 and 9.7 conform to this convention. Consistency in the analysis formula requires that the principal components also be multiplied by $(\lambda_m)^{1/2}$, with the result that the variance of each u_m increases to λ_m^2. A major advantage of this rescaling, however, is that the eigenvector elements are more directly interpretable in terms of the relationship between the principal components and the original data. Under this rescaling, each eigenvector element $e_{k,m}$ is numerically equal to the correlation $r_{u,z}$ between the mth principal component u_m and the kth standardized variable z_k. Also, this scaling is usually applied if a subset of the eigenvectors are to be rotated.

The last scaling shown in Table 9.4, resulting in $\lVert \mathbf{e}_m \rVert \equiv (\lambda_m)^{-1/2}$, is less commonly used. This scaling is achieved by dividing each element of the original unit-length eigenvectors by the square root of the corresponding eigenvalue. The resulting expression for the correlations between the principal components and the original data are more awkward, but this scaling has the advantage that all principal components have equal, unit variance. This property can be useful in the detection of outliers.

9.4 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a statistical technique that identifies a sequence of pairs of patterns in two multivariate data sets, and constructs sets of transformed variables by projecting the original data onto these patterns. The approach thus bears some similarity to PCA, which searches for patterns within a single multivariate data set that represent maximum amounts of the variation in

the data. In CCA, the patterns are chosen such that the new variables defined by projection of the two data sets onto these patterns exhibit maximum correlation, but are uncorrelated with the projections of the data onto any of the other identified patterns. That is, CCA identifies new variables that maximize the interrelationships between two data sets, in contrast to the patterns describing the internal variability within a single data set identified in PCA. It is this sense that CCA has been referred to as a "double-barreled PCA."

Canonical correlation analysis can also be viewed as an extension of multiple regression to the case of a vector-valued predictand variable $\mathbf{y}$. Ordinary multiple regression finds a weighted average, or pattern, of the vector of predictor variables $\mathbf{x}$ such that the correlation between the dot product $\mathbf{b}^{\mathrm{T}}\mathbf{x}$ and the scalar predictand y is maximized. Here the elements of the vector $\mathbf{b}$ are the ordinary least-squares regression coefficients computed using the methods described in Chapter 6, and $\mathbf{b}^{\mathrm{T}}\mathbf{x}$ is a new variable called the "predicted value of y," or $\hat{y}$. Canonical correlation analysis looks for pairs of sets of weights analogous to the regression coefficients, such that the correlations between the new variables defined by the respective dot products with $\mathbf{x}$ and (the vector) $\mathbf{y}$ are maximized.

As is also the case with PCA, CCA has been most widely applied to geophysical data in the form of fields. In this setting the vector $\mathbf{x}$ contains observations of one variable at a collection of gridpoints or locations, and the vector $\mathbf{y}$ contains observations of a different variable at a set of locations that may or may not be the same as those represented in $\mathbf{x}$. Typically the data consist of time series of observations of the two fields. When individual observations of the fields $\mathbf{x}$ and $\mathbf{y}$ are made simultaneously, a CCA can be useful in diagnosing aspects of the coupled variability of the two fields (e.g., Nicholls, 1987). When observations of $\mathbf{x}$ precede observations of $\mathbf{y}$ in time, the CCA may lead to statistical forecasts of the $\mathbf{y}$ field using the $\mathbf{x}$ field as a predictor (e.g., Barnston and Ropelewski, 1992). A more comprehensive comparison between CCA and PCA in the context of atmospheric data analysis is included in Bretherton *et al.* (1992).

9.4.1 Canonical Variates

A CCA transforms pairs of original data vectors $\mathbf{x}$ and $\mathbf{y}$ into sets of new variables, called *canonical variates*, v_m and w_m, defined by the dot products

$$v_m = \mathbf{a}_m^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{I} a_{m,i}x_i, \qquad m = 1, \ldots, \min(I, J); \tag{9.46a}$$

and

$$w_m = \mathbf{b}_m^{\mathrm{T}}\mathbf{y} = \sum_{j=1}^{J} b_{m,j}y_j, \qquad m = 1, \ldots, \min(I, J). \tag{9.46b}$$

This construction of the canonical variates is similar to that of the principal com-

ponents u_m [Eq. (9.36)], in that each is a linear combination (a sort of weighted average) of elements of the respective data vectors $\mathbf{x}$ and $\mathbf{y}$. Note that the data vectors and their corresponding vectors of weights, $\mathbf{a}_m$ and $\mathbf{b}_m$, need not have the same dimensions. The vectors $\mathbf{x}$ and $\mathbf{a}_m$ each have I elements, and the vectors $\mathbf{y}$ and $\mathbf{b}_m$ each have J elements. The number of pairs, M, of canonical variates that can be extracted from the two data sets is equal to the smaller of the dimensions of $\mathbf{x}$ and $\mathbf{y}$, i.e., $M = \min(I, J)$.

The properties of CCA are such that the canonical vectors $\mathbf{a}_m$ and $\mathbf{b}_m$ are chosen which result in the canonical variates having the properties

$$\mathrm{Corr}[v_1, w_1] \geq \mathrm{Corr}[v_2, w_2] \geq \cdots \geq \mathrm{Corr}[v_M, w_M] \geq 0; \tag{9.47a}$$

$$\mathrm{Corr}[v_k, w_m] = \begin{cases} r_{C_m}, & k = m \\ 0, & k \neq m \end{cases}; \tag{9.47b}$$

and

$$\mathrm{Var}[v_m] = \mathrm{Var}[w_m] = 1, \qquad m = 1, \ldots, M. \tag{9.47c}$$

Equation (9.47a) states that each of the M successive pairs of canonical variates exhibits a weaker correlation than the previous pair. These (Pearson product–moment) correlations between the pairs of canonical variates are called the *canonical correlations*, r_C. The canonical correlations can always be expressed as positive numbers, since either $\mathbf{a}_m$ or $\mathbf{b}_m$ can be multiplied by -1 if necessary. Equation (9.47b) states that each canonical variates is uncorrelated with all of the other canonical variates except its specific counterpart in the mth pair. Finally, Eq. (9.47c) states that each of the canonical variates has unit variance.

9.4.2 Computation of CCA

The information drawn upon by CCA is contained in the joint variance–covariance matrix of the variables $\mathbf{x}$ and $\mathbf{y}$. For purposes of computing this matrix, the two data vectors are concatenated into a single vector $\mathbf{c}^{\mathrm{T}} = [\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}}]$. This vector contains $I + J$ elements, the first I of which are the elements of $\mathbf{x}$, and the last J of which are the elements of $\mathbf{y}$. The $(I + J \times I + J)$ variance-covariance matrix of $\mathbf{c}$, $[\mathbf{S}_C]$, is then partitioned into four blocks, in a manner similar to that done for the correlation matrix in Fig. 9.6:

$$[\mathbf{S}_C] = \frac{1}{n-1}[\mathbf{C}']^{\mathrm{T}}[\mathbf{C}'] = \begin{bmatrix} \underset{(I\times I)}{[\mathbf{S}_{\mathbf{xx}}]} & \underset{(I\times J)}{[\mathbf{S}_{\mathbf{xy}}]} \\ \underset{(J\times I)}{[\mathbf{S}_{\mathbf{yx}}]} & \underset{(J\times J)}{[\mathbf{S}_{\mathbf{yy}}]} \end{bmatrix}. \tag{9.48}$$

Each of the n rows of the $(n \times I + J)$ matrix $[\mathbf{C}']$ contains one observation of the vector $\mathbf{x}'$ and one observation of the vector $\mathbf{y}'$, where as before the primes indicate centering of the data by subtraction of each of the respective sample means. The

$(I \times I)$ matrix $[\mathbf{S}_{xx}]$ is the variance–covariance matrix of the I variables in $\mathbf{x}$. The $(J \times J)$ matrix $[\mathbf{S}_{yy}]$ is the variance–covariance matrix of the J variables in $\mathbf{y}$. The matrices $[\mathbf{S}_{xy}]$ and $[\mathbf{S}_{yx}]$ contain the covariances between each of the elements of $\mathbf{x}$ and each element of $\mathbf{y}$, and they are related according to $[\mathbf{S}_{xy}] = [\mathbf{S}_{yx}]^T$.

The canonical correlations r_C are given by the square roots of the nonzero eigenvalues of the matrices

$$[\mathbf{M}_x] = [\mathbf{S}_{xx}]^{-1}[\mathbf{S}_{xy}][\mathbf{S}_{yy}]^{-1}[\mathbf{S}_{yx}] \tag{9.49a}$$

and

$$[\mathbf{M}_y] = [\mathbf{S}_{yy}]^{-1}[\mathbf{S}_{yx}][\mathbf{S}_{xx}]^{-1}[\mathbf{S}_{xy}]. \tag{9.49b}$$

The matrix $[\mathbf{M}_x]$ is dimensioned $(I \times I)$, and the matrix $[\mathbf{M}_y]$ is dimensioned $(J \times J)$. The first $M = \min(I, J)$ eigenvalues of these two matrices will be identical, and if $I \neq J$, the remaining eigenvalues of the larger matrix will all be zero. Furthermore, the canonical vectors $\mathbf{a}_m$ and $\mathbf{b}_m$ are the respective eigenvectors of these matrices, satisfying

$$[\mathbf{M}_x]\mathbf{a}_m = r^2_{C_m}\mathbf{a}_m, \qquad m = 1, \ldots, M; \tag{9.50a}$$

and

$$[\mathbf{M}_y]\mathbf{b}_m = r^2_{C_m}\mathbf{b}_m, \qquad m = 1, \ldots, M. \tag{9.50b}$$

Here r_{C_m} is the mth canonical correlation. In general the matrices $[\mathbf{M}_x]$ and $[\mathbf{M}_y]$ are not symmetric, so that finding their eigenvalues and eigenvectors is more demanding computationally than the corresponding calculations required for PCA.

Canonical correlation analysis can be performed on standardized variables as well, in which case the dispersion matrix in Eq. (9.48) will be the $(I + J \times I + J)$ joint correlation matrix of the standardized variables. Basing the CCA on this correlation matrix yields exactly the same canonical correlations, but will produce different values for the canonical vectors $\mathbf{a}_m$ and $\mathbf{b}_m$.

Example 9.7. CCA of the January 1987 Temperature Data

A simple illustration of the mechanics of a small CCA can be provided by again analyzing the January 1987 temperature data for Ithaca and Canandaigua, New York, given in Table A.1. Let the $I = 2$ Ithaca temperature variables be $\mathbf{x} = [T_{max}, T_{min}]^T$, and similarly let the $J = 2$ Canandaigua temperature variables be $\mathbf{y}$. The joint covariance matrix $[\mathbf{S}_C]$ of these quantities is then the (4×4) matrix

$$[\mathbf{S}_C] = \begin{bmatrix} 59.516 & 75.433 & 58.070 & 51.697 \\ 75.433 & 185.467 & 81.633 & 110.800 \\ 58.070 & 81.633 & 61.847 & 56.119 \\ 51.697 & 110.800 & 56.119 & 77.581 \end{bmatrix}. \tag{9.51}$$

This symmetric matrix contains the sample variances of the four variables on the diagonal, and the covariances between the variables in the other positions. This matrix is related to the corresponding elements of the correlation matrix involving the same variables (Table 3.3) through the square roots of the diagonal elements: Dividing each element by the square roots of the diagonal elements in its row and column produces the corresponding correlation matrix. This is the operation shown in Eq. (9.20).

Since $I = J$ in this example, the matrices required to compute $[\mathbf{M}_x]$ and $[\mathbf{M}_y]$ using Eq. (9.49) are obtained by "quartering" $[\mathbf{S}_C]$, to yield

$$[\mathbf{S}_{xx}] = \begin{bmatrix} 59.516 & 75.433 \\ 75.433 & 185.467 \end{bmatrix}, \tag{9.52a}$$

$$[\mathbf{S}_{yy}] = \begin{bmatrix} 61.847 & 56.119 \\ 56.119 & 77.581 \end{bmatrix}, \tag{9.52b}$$

and

$$[\mathbf{S}_{yx}] = \begin{bmatrix} 58.070 & 81.633 \\ 51.697 & 110.800 \end{bmatrix}, \tag{9.52c}$$

with the fourth (upper-right) matrix given by the transpose of Eq. (9.52c): $[\mathbf{S}_{xy}] = [\mathbf{S}_{yx}]^T$. Applying Eq. (9.49) yields

$$[\mathbf{M}_x] = \begin{bmatrix} .8304 & .3771 \\ .0682 & .7004 \end{bmatrix} \tag{9.53a}$$

and

$$[\mathbf{M}_y] = \begin{bmatrix} .8452 & .2589 \\ .0910 & .6856 \end{bmatrix}. \tag{9.53b}$$

The $M = 2$ eigenvectors of $[\mathbf{M}_x]$, $\mathbf{a}_1$ and $\mathbf{a}_2$, and the corresponding eigenvectors of $[\mathbf{M}_y]$, $\mathbf{b}_1$ and $\mathbf{b}_2$, are shown in Table 9.5. The first element of each pertains to the respective maximum temperature variable, and the second elements pertain to the minimum temperature variables. The first eigenvalue of these two matrices is $\lambda_1 = 0.5020$, yielding the first canonical correlation as $\sqrt{\lambda_1} = 0.709$. The second canonical correlation of 0.038 is so small that this canonical pair would not be of practical interest.

The time series of the first pair of canonical variables is given by the dot products of $\mathbf{a}_1$ and $\mathbf{b}_1$ with the pairs of temperature values for Ithaca and Canandaigua, respectively, from Table A.1. The value of v_1 for January 1 would be constructed as (33)(.9613) + (19)(.2753) = 36.95. The the time series of v_1 (pertaining to the Ithaca temperatures) would consist of 31 values (one for each day): 36.95, 37.46, 34.90, 27.60, 25.13, . . . , 31.87, 39.02. Similarly, the time series for w_1 (pertaining to the Canandaigua temperatures) is 41.47, 43.35, 37.03, 33.72, 33.64, . . . , 33.91,

Table 9.5

Canonical Vectors $\mathbf{a}_m$ (Corresponding to Ithaca Temperatures) and $\mathbf{b}_m$ (Corresponding to Canandaigua Temperatures) for Partition of Covariance Matrix in Eq. (9.51) with $I = J = 2^a$

	$\mathbf{a}_1$ (Ithaca)	$\mathbf{b}_1$ (Canandaigua)	$\mathbf{a}_2$ (Ithaca)	$\mathbf{b}_2$ (Canandaigua)
$T_{\max}$	.9613	.9409	−.8457	−.7155
$T_{\min}$	.2753	.3386	−.7155	.6986
λ_m	0.5020		0.00139	
$r_{C_m} = \sqrt{\lambda_m}$	0.709		0.038	

[a] These vectors are the eigenvectors of the matrices in Eq. (9.53). Also shown are the eigenvalues λ_m of $[\mathbf{M}_x]$ and $[\mathbf{M}_y]$, and the canonical correlations which are their square roots.

43.54. The first canonical correlation coefficient, $r_{C_m} = 0.709$, is the correlation between this first pair of canonical variables, v_1 and w_1, and is the largest possible correlation between pairs of linear combinations of these two data sets. ¤

9.4.3 CCA Applied to Fields

Canonical correlation analysis is usually most interesting for atmospheric data when applied to fields. Here the spatially distributed observations (either at grid-points or observing locations) are encoded into the vectors **x** and **y** in the same way as for PCA. That is, even though the data are drawn from a two- or three-dimensional field, each location is numbered sequentially and pertains to one element of the corresponding data vector. It is not necessary for the spatial domains encoded into **x** and **y** to be the same, and, indeed, in the applications of CCA that have appeared in the literature they are usually different.

As is the case with the use of PCA with spatial data, it is often informative to plot maps of the canonical vectors by associating the magnitudes of their elements and the geographic locations to which they pertain. In this context the canonical vectors are sometimes called *canonical patterns*, since the resulting maps show spatial patterns of the ways in which the original variables contribute to the canonical variables. Examining the pairs of maps formed by corresponding vectors $\mathbf{a}_m$ and $\mathbf{b}_m$ can be informative about the nature of the relationship between variations in the data over the two domains encoded in **x** and **y**, respectively.

Figure 9.9, from Wallace *et al.* (1992), shows one such pair of canonical patterns. Figure 9.9a shows the spatial distribution of the elements of a canonical vector **a**, pertaining to a data vector **x** that contains values of average December–January sea–surface temperatures (SSTs) in the north Pacific Ocean. This pattern accounts for 18% of the total variance of the SSTs in the data set analyzed. Figure 9.9b shows the spatial distribution of the elements of the corresponding canonical vector **b**, which pertains to average hemispheric 500-mb heights **y** for

Fig. 9.9 Spatial displays of a pair of canonical vectors pertaining to average winter sea–surface temperatures (SSTs) in the northern Pacific Ocean (a), and hemispheric winter 500-mb heights (b). The pattern of SST anomalies in panel a (and its negative) are associated with the PNA pattern of 500-mb height anomalies shown in panel b. The correlation between SSTs projected onto panel a and 500-mb heights projected onto panel b is the canonical correlation 0.79. [From Wallace *et al.* (1992), reproduced with permission of the American Meteorological Society.]

the same winters included in the SST data in **x**. This pattern accounts for 23% of the total variance of the winter hemispheric height variations. The sea-surface temperature pattern described by the canonical vector on the left is one of either cold water in the central north Pacific and warm water along the west coast of North America, or warm water in the central north Pacific and cold water along the west coast of North America. The pattern of 500-mb height variations in Fig. 9.8b is remarkably similar to the PNA pattern (cf. Figs. 9.8b and 3.20). Taken together, these figures indicate that cold SSTs in the central Pacific simultaneously with warm SSTs in the northeast Pacific tend to coincide with a 500-mb ridge over northwestern North America and a 500-mb trough over southeastern North America. Similarly, warm water in the central north Pacific and cold water in the northwestern Pacific are associated with the more zonal PNA flow. The time series of the canonical variable v is obtained by projecting (i.e., forming the dot products of) the time series of SSTs **x** onto the vector **a** shown in Fig. 9.9a, and the time series of the canonical variable w results from the 500 mb heights **y** being projected onto the vector **b** shown in Fig. 9.9b. The correlation between the two time series v and w is the canonical correlation $r_C = 0.79$.

Figure 9.10, from Barnett and Preisendorfer (1987), illustrates the application of CCA to atmospheric fields where the observations **x** and **y** have not been made simultaneously. Here **y** is average January surface temperature over the United States, and **x** is average sea-level pressures over much of the Northern Hemisphere, during the preceding October, November, and December. The sea-level pressure fields each consisted of 280 gridpoints, so each vector **x** consists of $I = (3)(280) = 840$ elements. Here, the first 280 elements of **x** contain the October pressures, the second 280 elements contain the November pressures, and the third group of 280 elements contain the December pressures. Essentially the same approach to simultaneous consideration of multiple fields has been taken as in the PCA results shown in Fig. 9.7, with data from several fields being concatenated into a single vector.

However, the essential feature of the construction of the CCA shown in Fig. 9.10 is that the **x** fields all precede the **y** field, so that, at least in principle, the CCA can be used to forecast **y**, if **x** is known. The forecast is constructed using the pattern vector shown in Fig. 9.10a to forecast the pattern in Fig. 9.10b. That is, these maps indicate that there is a tendency for the pressure anomaly patterns shown in Fig. 9.10a to be followed by the temperature patterns shown in Fig. 9.10b. As before, these maps actually specify two sets of patterns, depending on which sign is applied. They indicate that low December pressures in the north Pacific and western Atlantic, with high December pressures over western North America (with similar but weaker patterns during November and October) have a tendency to be followed by relatively warm January temperatures in western North America and relatively cool January temperatures in eastern North America. In years when December pressures are relatively high over the north Pacific

a
g_1 – MAP SLP PREDICTING JANUARY
180°
0°
90°N
0°
40
H
-20
L
L
-40
H
L
-40
H
-20
OCTOBER SLP
90°N
0°
20
H
H
20
-20
L
L
-20
NOVEMBER SLP
90°N
0°
H
20
L
-60
H
40
L
-60
-20
-20
-20
DECEMBER SLP

h_1 – MAP SLP PREDICTION JANUARY
b
0.3
0
-0.3
0.4
-0.5

and western Atlantic, and are relatively low over western North America, these maps indicate that North American temperatures in the following January tend to be relatively cool in the west and relatively warm in the east.

When using CCA in an exploratory or diagnostic analysis, it may not be of interest to reconstruct contributions to the original data made by a set of canonical patterns. However, when used as part of a forecasting procedure, it is necessary operationally to reconstruct estimated data values predicting the **y** field using some number of the canonical vectors $\mathbf{b}_m$. Usually it is more convenient to work with the fields of anomalies, $\mathbf{x}'$ and $\mathbf{y}'$, but note that the covariance matrices for the anomalies ($[\mathbf{S}_{x'x'}]$ and $[\mathbf{S}_{y'y'}]$) and the covariance matrices for the original data ($[\mathbf{S}_{xx}]$ and $[\mathbf{S}_{yy}]$) are identical. This "synthesis" process is in general more difficult than the corresponding procedure in PCA [Eqs. (9.38) and (9.39)], because the canonical vectors are not orthogonal. However, a synthesis equation can be derived for CCA by manipulating the matrix "analysis" equation

$$\underset{(n \times M^*)}{[\mathbf{W}]} = \underset{(n \times J)}{[\mathbf{Y}']} \underset{(J \times M^*)}{[\mathbf{B}]}. \tag{9.54}$$

Here $[\mathbf{Y}']$ is the full matrix of predictand anomaly values; each of the n rows of this matrix contains one data record $\mathbf{y}'^{\mathrm{T}}$. The matrix $[\mathbf{B}]$ has as its columns some number M^* (not necessarily the full number M) of the canonical vectors **b** pertaining to $\mathbf{y}'$, and $[\mathbf{W}]$ is the matrix of the corresponding M^* canonical scores. Equation (9.54) can be solved for $[\mathbf{Y}']$ by multiplying on the right by $[\mathbf{B}]^{\mathrm{T}}$, and then multiplying on the right again by $([\mathbf{B}][\mathbf{B}]^{\mathrm{T}})^{-1}$:

$$[\mathbf{W}][\mathbf{B}]^{\mathrm{T}} = [\mathbf{Y}'][\mathbf{B}][\mathbf{B}]^{\mathrm{T}} \tag{9.55a}$$

$$\begin{aligned} [\mathbf{W}][\mathbf{B}]^{\mathrm{T}}([\mathbf{B}][\mathbf{B}]^{\mathrm{T}})^{-1} &= [\mathbf{Y}'][\mathbf{B}][\mathbf{B}]^{\mathrm{T}}([\mathbf{B}][\mathbf{B}]^{\mathrm{T}})^{-1} \\ &= [\mathbf{Y}'][\mathbf{I}] \\ &= [\mathbf{Y}']. \end{aligned} \tag{9.55b}$$

The relationship between $[\mathbf{V}]$, $[\mathbf{X}']$, and $[\mathbf{A}]$ is analogous. The forecast quantities are recovered from the anomalies simply by adding back the appropriate mean value. If the CCA has been computed using standardized variables, so that Eq. (9.48) is a correlation matrix, the dimensional values of the forecast variables need to be reconstructed by multiplying by the appropriate standard deviation and adding the appropriate mean [i.e., by reversing Eq. (3.16) or (4.20) to yield Eq. (4.22)].

Fig. 9.10 Spatial displays of a pair of canonical vectors pertaining to average sea-level pressures (SLPs) in October, November, and December (a) and average temperatures over the United States during the following January (b). Since the sea-level pressures in **x** are observed prior to the surface temperatures **y**, the link between the two fields made by the CCA may be useful for temperature forecasts. [From Barnett and Preisendorfer (1987), reproduced with permission of the American Meteorological Society.]

9.4.4 Combining PCA and CCA

In practical applications it is sometimes useful to "prefilter" the two fields of raw data using PCA before subjecting them to a CCA. Thus, rather than directly correlating the fields $\mathbf{x}$ and $\mathbf{y}$, the combined analysis operates on the fields $\mathbf{u}_x$ and $\mathbf{u}_y$, which consist of the first principal components of $\mathbf{x}$ and $\mathbf{y}$. That is, separate PCAs are computed for each of the two fields, and these two analyses may be truncated at different numbers of principal components M_x^* and M_y^*.

This combined approach is not always best, and can be inferior if important information is discarded when truncating the PCA. However, combining CCA with PCA is attractive in circumstances where there are large numbers I and/or J of observations on the original variables, especially if the $\mathbf{x}$ and $\mathbf{y}$ fields are strongly spatially correlated. Using a PCA to reduce the dimensionality of these original fields reduces the number of computations necessary to carry out the CCA while retaining most of the original information. In instances where the sample size, n, is not large compared to I and J, preprocessing the data with a PCA also tends to improve the stability (i.e., sampling properties) of the CCA. This stability consideration is especially important when a forecasting application is contemplated, since good performance on independent data is required.

9.5 Discriminant Analysis

The final two sections in this chapter deal with classification of multivariate data. Given a collection of data vectors $\mathbf{x}$ consisting of K elements, it is desired to separate these into groups on the basis of the individual vector elements x_k. If the groups are specified in advance, the classification problem is one of discriminating among the known groups. Here it is necessary to have available a "training sample" in which it is known that each of the vectors is correctly classified, in order to build rules according to which future observations $\mathbf{x}$ of unknown group origin can be similarly classified. This is the problem of discriminant analysis, presented in this section.

If the criterion according to which the observations $\mathbf{x}$ have been grouped pertains to a time after the observation of $\mathbf{x}$, then the discriminant analysis can be employed as a forecasting procedure. Here each group represents a discrete outcome, and the forecast consists of a categorical statement that one of these outcomes will occur. The forecast is made by "classifying" the current observation $\mathbf{x}$ as belonging to the group that is forecast to occur. Examples of the use of discriminant analysis for forecasting can be found in Lawson and Cerveny (1985), Miller (1962), and Ward and Folland (1991).

For some data sets it is not known to which group each observation belongs, or possibly even how many groups exist. In this case discriminant analysis cannot be applied. However, degrees of similarity and difference among different vector observations $\mathbf{x}$ can be used to construct groups on the basis only of the data at hand.

That is, one looks to see how the data points "clump" or cluster in their K-dimensional space. This is the problem of cluster analysis, presented in Section 9.6.

9.5.1 Linear Discriminant Analysis

The simplest form of discriminant analysis involves distinguishing between two groups on the basis of a K-dimensional vector of observations $\mathbf{x}$. A training sample must exist, consisting of n_1 observations of $\mathbf{x}$ known to have come from group 1, and n_2 observations of $\mathbf{x}$ known to have come from group 2. That is, the basic data are two data matrices, $[\mathbf{X}_1]$, dimensioned $(n_1 \times K)$, and $[\mathbf{X}_2]$, dimensioned $(n_2 \times K)$. The goal is to find a linear function of the K elements of the observation vector, called the *discriminant function*, that will best allow a future K-dimensional vector of observations $\mathbf{y}$ to be classified as belonging to either group 1 or group 2.

Each of the two groups is characterized by a K-dimensional mean vector

$$\bar{\mathbf{x}}_g = \frac{1}{n_g}[\mathbf{X}_g]^{\mathrm{T}}\mathbf{1} = \begin{bmatrix} \frac{1}{n_g}\sum_{i=1}^{n_g} x_{i,1} \\ \frac{1}{n_g}\sum_{i=1}^{n_g} x_{i,2} \\ \vdots \\ \frac{1}{n_g}\sum_{i=1}^{n_g} x_{i,K} \end{bmatrix}, \qquad g = 1, 2; \tag{9.56}$$

where $\mathbf{1}$ is a $(n \times 1)$ vector containing only 1's. That is, the K elements of each of the two mean vectors are computed by averaging each of the K elements of the data vectors $\mathbf{x}$. The averaging in Eq. (9.56) is done separately for the observations belonging to groups 1 and 2.

Nearly always in discriminant analysis, the assumption is made that the populations underlying each of the groups has the same covariance matrix. That is, the mean vectors characterizing each group may be different (discrimination between groups would be very difficult if they were not), but the nature of the dispersion of the data points around these mean vectors is assumed to be the same for each group. The sample covariance matrices, $[\mathbf{S}_1]$ and $[\mathbf{S}_2]$, are computed from the data matrices, $[\mathbf{X}_1]$ and $[\mathbf{X}_2]$, using Eqs. (9.18) and (9.19). Since the covariance structures are assumed to be equal, these two variance–covariance matrices are averaged to yield a pooled estimate of the dispersion of the data around their means,

$$[\mathbf{S}_{\text{pooled}}] = \frac{(n_1 - 1)[\mathbf{S}_1] + (n_2 - 1)[\mathbf{S}_2]}{(n_1 + n_2 - 2)}. \tag{9.57}$$

If $n_1 = n_2$, Eq. (9.57) results in each element of $[\mathbf{S}_{\text{pooled}}]$ being constructed as the simple average of the corresponding elements of $[\mathbf{S}_1]$ and $[\mathbf{S}_2]$.

The fundamental idea behind linear discriminant analysis is to find a direction $\mathbf{d}_1$ in the K-dimensional space of the data, such that the distance between the two mean vectors is maximized when the data are projected onto $\mathbf{d}_1$. Finding this direction reduces the discrimination problem from one of sifting through the relationships among the K elements of the data vectors to looking at a single number. That is, the data vector $\mathbf{x}$ is transformed to a new variable, $\delta_1 = \mathbf{d}_1^T\mathbf{x}$, known as *Fisher's linear discriminant function*. The groups of K-dimensional multivariate data are essentially reduced to groups of univariate data with different means, distributed along the $\mathbf{d}_1$ axis. The discriminant vector locating this direction of maximum separation is given by

$$\mathbf{d}_1 = [\mathbf{S}_{\text{pooled}}]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{9.58a}$$

or, if a unit vector is more convenient,

$$\mathbf{d}_1 = \frac{[\mathbf{S}_{\text{pooled}}]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\|[\mathbf{S}_{\text{pooled}}]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\|}. \tag{9.58b}$$

Either of the scalings in Eq. (9.58), or any other constant multiple of $\mathbf{d}_1$, can be used equally well, since all of these vectors will point in the same direction.

The decision to classify a future observation $\mathbf{y}$ as belonging to either group 1 or group 2 can now be made according to the value of the scalar $\delta_1 = \mathbf{d}_1^T\mathbf{y}$. This dot product is a one-dimensional (i.e., scalar) projection of the vector $\mathbf{y}$ onto the direction of maximum separation, $\mathbf{d}_1$. The discriminant function δ_1 is essentially a new variable, analogous to the new variable u in PCA and the new variables v and w in CCA, produced as a linear combination of the elements of a data vector $\mathbf{y}$. In the simplest application, the observation $\mathbf{y}$ is classified as belonging to group 1 if the projection $\mathbf{d}_1^T\mathbf{y}$ is closer to the projection of the group 1 mean onto the direction $\mathbf{d}_1$, and is classified as belonging to group 2 if $\mathbf{d}_1^T\mathbf{y}$ is closer to the projection of the mean of group 2.

Along the $\mathbf{d}_1$ axis, the midpoint between the means of the two groups is given by the projection of the average of these two mean vectors onto the vector $\mathbf{d}_1$:

$$\bar{\delta}_1 = \mathbf{d}_1^T\frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}. \tag{9.59}$$

This value defines the dividing line between values of the discriminant function $\mathbf{d}_1^T\mathbf{y}$ for a future observation $\mathbf{y}$ that would result in its being assigned to either group 1 or group 2. That is, a future observation $\mathbf{y}$ whose group membership is unknown can be assigned to a group according to the rule

$$\text{Assign } \mathbf{y} \text{ to group 1 if} \qquad \mathbf{d}_1^T\mathbf{y} - \bar{\delta}_1 \geq 0, \tag{9.60a}$$

or

$$\text{Assign } \mathbf{y} \text{ to group 2 if} \qquad \mathbf{d}_1^T\mathbf{y} - \bar{\delta}_1 < 0. \tag{9.60b}$$

Example 9.8. A Linear Discriminant Analysis for $K = 2$-Dimensional Data

Table 9.6 shows average July temperature and precipitation for stations in three regions of the United States. The data vectors are composed of $K = 2$ elements each: one temperature element and one precipitation element. Consider the problem of distinguishing between membership in group 1 and group 2. This problem might arise if the stations in Table 9.6 represented the "core" portions of their respective climatic regions, and on the basis of these data one wanted to classify stations not listed in this table as belonging to one or the other of these two groups.

The mean vectors for the $n_1 = 10$ and $n_2 = 9$ data vectors in groups 1 and 2 are

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 80.6°\text{F} \\ 5.67 \text{ in.} \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 78.7°\text{F} \\ 3.57 \text{ in.} \end{bmatrix}, \tag{9.61a}$$

and the two sample variance–covariance matrices are

$$[\mathbf{S}_1] = \begin{bmatrix} 1.47 & 0.65 \\ 0.65 & 1.45 \end{bmatrix} \quad \text{and} \quad [\mathbf{S}_2] = \begin{bmatrix} 2.08 & 0.06 \\ 0.06 & 0.17 \end{bmatrix}. \tag{9.61b}$$

Since $n_1 \neq n_2$, the pooled estimate for the common variance–covariance matrix is obtained by the weighted average specified by Eq. (9.57). The direction $\mathbf{d}_1$ pointing in the direction of maximum separation of the two sample mean vectors is then computed using Eq. (9.58a) as

$$\begin{aligned} \mathbf{d}_1 &= \begin{bmatrix} 1.76 & 0.37 \\ 0.37 & 0.85 \end{bmatrix}^{-1} \left(\begin{bmatrix} 80.6 \\ 5.67 \end{bmatrix} - \begin{bmatrix} 78.7 \\ 3.57 \end{bmatrix} \right) \\ &= \begin{bmatrix} 0.625 & -0.272 \\ -0.272 & 1.295 \end{bmatrix} \begin{bmatrix} 1.9 \\ 2.10 \end{bmatrix} = \begin{bmatrix} 0.62 \\ 2.20 \end{bmatrix}. \end{aligned} \tag{9.62}$$

Figure 9.11 illustrates the geometry of this problem. Here the data for the warmer and wetter southeastern stations of group 1 are plotted as circles, and the central U.S. stations of group 2 are plotted as X's. The vector means for the two groups are also indicated by the heavy symbols. The projections of these two means onto the direction $\mathbf{d}_1$ are indicated by the lighter dashed lines. The midpoint between these two projections locates the dividing point between the two groups in the one-dimensional discriminant space defined by $\mathbf{d}_1$. The heavy dashed line perpendicular to the discriminant function at this point divides the (temperature, precipitation) plane into two regions. Future points $\mathbf{y}$ of unknown group membership falling above and to the right of this heavy line will be classified as belonging to group 1, and points falling below and to the left will be classified as belonging to group 2.

Since the average of the mean vectors for groups 1 and 2 is $[79.65, 4.62]^T$, the value of the dividing point is $= (0.62)(79.65) + (2.20)(4.62) = 59.55$. Of the 19 points in this training data, only that for Atlanta has been misclassified. For this

Table 9.6

Average July Temperature (Temp., °F) and Precipitation (Ppt., in.) for Locations in Three Regions of the United States[a]

Group 1: southeastern USA			Group 2: central USA			Group 3: northeastern USA		
Station	Temp	Ppt	Station	Temp	Ppt	Station	Temp	Ppt
Athens, GA	79.2	5.18	Concordia, KS	79.0	3.37	Albany, NY	71.4	3.00
Atlanta, GA	78.6	4.73	Des Moines, IA	76.3	3.22	Binghamton, NY	68.9	3.48
Augusta, GA	80.6	4.40	Dodge City, KS	80.0	3.08	Boston, MA	73.5	2.68
Gainesville, FL	80.8	6.99	Kansas City, MO	78.5	4.35	Bridgeport, CT	74.0	3.46
Huntsville, AL	79.3	5.05	Lincoln, NE	77.6	3.20	Burlington, VT	69.6	3.43
Jacksonville, FL	81.3	6.54	Springfield, MO	78.0	3.58	Hartford, CT	73.4	3.09
Macon, GA	81.4	4.46	St. Louis, MO	78.9	3.63	Portland, ME	68.1	2.83
Montgomery, AL	81.7	4.78	Topeka, KS	78.6	4.04	Providence, RI	72.5	3.01
Pensacola, FL	82.3	7.18	Wichita, KS	81.4	3.62	Worcester, MA	69.9	3.58
Savannah, GA	81.2	7.37						
Averages	80.6	5.67		78.7	3.57		71.3	3.17

[a] Averages are for the period 1951–1980. [From Quayle and Presnell (1991).]

Fig. 9.11 Geometry of linear discriminant analysis applied to the southeastern (circles) and central (Xs) U.S. data in Table 9.6. The (vector) means of the two groups of data are indicated by the heavy symbols, and their projections onto the discriminant function are indicated by the light dashed arrows. The midpoint between these two projections, $\bar{\delta}_1$, defines the dividing line (heavier dashed line) used to assign future (temperature, precipitation) pairs to the groups. Of these training data, only the data point for Atlanta would be misclassified. Note that the discriminant function has been shifted to the right (i.e., does not pass through the origin) in order to improve the clarity of the plot, but that this does not affect the relative positions of the projections of the data points onto it.

station, $\delta_1 = \mathbf{d}_1^T\mathbf{x} = (0.62)(78.6) + (2.20)(4.73) = 59.14$. Since this value of δ_1 is slightly less than the midpoint value, Atlanta would be falsely classified as belonging to group 2 [cf. Eq. (9.60)]. By contrast, the point for Augusta lies just to the "group 1 side" of the heavy dashed line. For Augusta, $\delta_1 = \mathbf{d}_1^T\mathbf{x} = (0.62)(80.6) + (2.20)(4.40) = 59.65$, which is slightly greater than the cutoff value.

Consider now the assignment of two stations not listed in Table 9.6 to either group 1 or group 2. For New Orleans, LA the average July temperature is 82.1°F, and the average July precipitation is 6.73 in. Applying Eq. (9.60), one finds $\mathbf{d}_1^T\mathbf{y} - \delta_1 = (0.62)(82.1) + (2.20)(6.73) - 59.55 = 65.71 - 59.55 = 6.16 > 0$. Therefore, New Orleans would be classified as belonging to group 1. Similarly,

the average July temperature and precipitation for Columbus, OH are 74.7°F and 3.37 in., respectively. For this station, $\mathbf{d}_1^T\mathbf{y} - \delta_1 = (0.62)(74.7) + (2.20)(3.37) - 59.55 = 53.73 - 59.55 = -5.82 < 0$, which would result in Columbus being classified as belonging to group 2. ¤

Example 9.8 was constructed with $K = 2$ observations in each data vector in order to allow the geometry of the problem to be represented in two dimensions. However, it is not necessary to restrict the use of discriminant analysis to situations with only bivariate observations. In fact, discriminant analysis is potentially most powerful when allowed to operate on higher-dimensional data. For example, it would be possible to extend Example 9.8 to classifying stations according to average temperature and precipitation for all 12 months. If this were done, each data vector $\mathbf{x}$ would consist of $K = 24$ values. However, the discriminant vector $\mathbf{d}_1$ would also consist of $K = 24$ elements, and the dot product $\delta_1 = \mathbf{d}_1^T\mathbf{y}$ would still be a single scalar that could be used to classify the group membership of $\mathbf{y}$.

Usually high-dimensional data vectors of atmospheric data exhibit substantial correlation among the K elements, and thus carry some redundant information. For example, the 12 monthly mean temperatures and 12 monthly mean precipitation values for a location do not provide 24 independent pieces of information. If only for computational economy, it can be a good idea to reduce the dimensionality of this kind of data before subjecting it to a discriminant analysis. This reduction in dimension is most commonly achieved through a principal-component analysis (Section 9.3). Since the groups in discriminant analysis are assumed to have the same covariance structure, it is natural to perform the PCA on the estimate of their common variance–covariance matrix, $[\mathbf{S}_{\text{pooled}}]$. If the data vectors are not of consistent units (e.g., some temperatures and some precipitation amounts), it will make more sense to perform the PCA on the corresponding correlation matrix. The discriminant analysis can then be carried out using M^*-dimensional data vectors containing elements that are the first M^* principal components, rather than the original K-dimensional raw data vectors. The resulting discriminant function will then pertain to the principal components, $\mathbf{u}$, rather than to the original data, $\mathbf{x}$. In addition, if the first two principal components account for a large fraction of the total variance, the data can be visualized in a plot such as that in Fig. 9.11, where the horizontal and vertical axes are the first two principal components.

The point on the discriminant function between the projections of the two sample means is not always the best point at which to make the separation between groups. One might have prior information that the probability of membership in group 1 is higher than that for group 2, perhaps because group 2 members are rather rare overall. If this is so, it would usually be desirable to move the classification boundary toward the group 2 mean, with the result that more future observations $\mathbf{y}$ would be classified as belonging to group 1. Similarly, if misclassifying a group 1 data value as belonging to group 2 were to be a more serious error than

misclassifying a group 2 data value as belonging to group 1, one would again want to move the boundary toward the group 2 mean. If p_1 is the prior probability (the probability according to previous information) that the observation $\mathbf{y}$ belongs to group 1, p_2 is the prior probability that the observation $\mathbf{y}$ belongs to group 2, $L(1 \mid 2)$ is the loss, or penalty, incurred when a group 2 member is incorrectly classified as part of group 2, and $L(2 \mid 1)$ is the loss incurred when a group 2 member is incorrectly classified as part of group 2, the classification rule in Eq. (9.60) can be revised to yield

$$\text{Assign } \mathbf{y} \text{ to group 1} \quad \text{if} \quad \mathbf{d}_1^{\mathrm{T}}\mathbf{y} - \bar{\delta}_1 \geq \ln\left[\frac{L(1 \mid 2)}{L(2 \mid 1)}\frac{p_2}{p_1}\right], \tag{9.63a}$$

or

$$\text{Assign } \mathbf{y} \text{ to group 2} \quad \text{if} \quad \mathbf{d}_1^{\mathrm{T}}\mathbf{y} - \bar{\delta}_1 < \ln\left[\frac{L(1 \mid 2)}{L(2 \mid 1)}\frac{p_2}{p_1}\right]. \tag{9.63b}$$

Note that if the misclassification losses are equal, the adjustment involves only the prior probabilities. If the prior probabilities are equal, the adjustment involves only the losses. If both the losses and the prior probabilities are equal, Eq. (9.63) reduces to Eq. (9.60).

9.5.2 Multiple Discriminant Analysis

The generalization of linear discriminant analysis to the case of more than two groups is called *multiple discriminant analysis.* Here the basic problem is to allocate a K-dimensional data vector $\mathbf{y}$ to one of $G > 2$ groups on the basis of one or more discriminant functions. The discriminant functions are computed on the basis of a training set of G data matrices $[\mathbf{X}_1], [\mathbf{X}_2], [\mathbf{X}_3], \ldots, [\mathbf{X}_G]$ dimensioned, respectively, $(n_g \times K)$. A sample variance-covariance matrix can be computed from each of the G sets of data, $[\mathbf{S}_1], [\mathbf{S}_2], [\mathbf{S}_3], \ldots, [\mathbf{S}_G]$, according to Eq. (9.19).

As in the two-group linear discriminant analysis, it is commonly assumed that the G groups are drawn from populations with different mean vectors, but having the same variance–covariance matrix. This common variance–covariance matrix is estimated by the weighted-average pooled estimate

$$[\mathbf{S}_{\text{pooled}}] = \frac{1}{n - G}\sum_{g=1}^{G}(n_g - 1)[\mathbf{S}_g], \tag{9.64}$$

where there are n_g observations in each group, and the total sample size is

$$n = \sum_{g=1}^{G} n_g. \tag{9.65}$$

The estimated common variance–covariance matrix in Eq. (9.64) is sometimes

called the *within-groups variance matrix*. Equation (9.57) is a special case of Eq. (9.64), with $G = 2$.

Computation of multiple discriminant functions also requires calculation of the *between-groups variance matrix*

$$[\mathbf{S}_\mathrm{B}] = \frac{1}{G-1}\sum_{g=1}^{G}(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_.)(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_.)^\mathrm{T}, \tag{9.66}$$

where

$$\bar{\mathbf{x}}_g = \frac{1}{n_g}[\mathbf{X}_g]^\mathrm{T}\mathbf{1} \tag{9.67a}$$

is the sample mean vector of the gth group, and

$$\bar{\mathbf{x}}_. = \frac{1}{n}\sum_{g=1}^{G} n_g\bar{\mathbf{x}}_g \tag{9.67b}$$

is the grand, or overall vector mean of all n observations. The matrix $[\mathbf{S}_\mathrm{B}]$ is essentially a variance–covariance matrix describing the dispersion of the G sample means around the overall mean [cf. Eq. (9.19)].

The number of discriminant functions that can be computed is the smaller of $G - 1$ and K. Thus for the two-group case discussed in the previous section, there is only $G - 1 = 1$ discriminant function, regardless of the dimensionality K of the data vectors. In the more general case, the discriminant functions are derived from the eigenvectors of the $(K \times K)$ matrix

$$[\mathbf{D}] = [\mathbf{S}_\mathrm{pooled}]^{-1}[\mathbf{S}_\mathrm{B}] \tag{9.68}$$

corresponding to the $M = \min(G - 1, K)$ nonzero eigenvalues. Usually computer routines for calculating eigenvectors will scale the results to unit length (i.e., $\|\mathbf{e}_m\| = 1$), whereas the discriminant vectors $\mathbf{d}_m$ are conventionally scaled such that

$$\mathbf{d}_m^\mathrm{T}[\mathbf{S}_\mathrm{pooled}]\mathbf{d}_m = 1. \tag{9.69}$$

This scaling is achieved by multiplying the unit-length eigenvectors $\mathbf{e}_m$ by constants c_m,

$$\mathbf{d}_m = c_m\mathbf{e}_m, \tag{9.70a}$$

where

$$c_m = (\mathbf{e}_m^\mathrm{T}[\mathbf{S}_\mathrm{pooled}]\mathbf{e}_m)^{-1/2}. \tag{9.70b}$$

The first discriminant function is then the dot product $\delta_1 = \mathbf{d}_1^\mathrm{T}\mathbf{x}$, the second discriminant function is $\delta_2 = \mathbf{d}_2^\mathrm{T}\mathbf{x}, \ldots$, and the Mth discriminant function is $\delta_M = \mathbf{d}_M^\mathrm{T}\mathbf{x}$. The M eigenvectors $\mathbf{d}_m$ define an M-dimensional *discriminant space*, in which the G groups of data exhibit maximum separation. The projections δ_m of the data onto these vectors are sometimes called the *discriminant coordinates*

or *canonical variates*. This second appellation is unfortunate and a cause of confusion, since these do not pertain to canonical correlation analysis.

As was also the case when distinguishing between $G = 2$ groups, observations $\mathbf{y}$ can be assigned to groups according to which of the G group means is closest in discriminant space. For the $G = 2$ case the discriminant space is one-dimensional, consisting only of a line. The group assignment rule in Eq. (9.60) is then particularly simple. More generally, it is necessary to evaluate the Euclidean distances between the candidate vector $\mathbf{y}$ and each of the G group means in order to find which is closest. It is actually easier to evaluate these in terms of squared distances, yielding the classification rule

Assign $\mathbf{y}$ to group g if

$$\sum_{m=1}^{M} [\mathbf{d}_m(\mathbf{y} - \bar{\mathbf{x}}_g)]^2 \leq \sum_{m=1}^{M} [\mathbf{d}_m(\mathbf{y} - \bar{\mathbf{x}}_h)]^2, \qquad \text{for all } h \neq g. \tag{9.71}$$

That is, the sum of the squared distances between $\mathbf{y}$ and each of the group means, along the directions defined by the vectors $\mathbf{d}_m$, are compared in order to find the closest group mean.

Example 9.9. Multiple Discriminant Analysis with G = *3 Groups*

Consider discriminating among all three groups of data in Table 9.6. Using Eq. (9.64), the pooled estimate of the common variance-covariance matrix is

$$[\mathbf{S}_{\text{pooled}}] = \frac{1}{28-3}\left(9\begin{bmatrix} 1.47 & 0.65 \\ 0.65 & 1.45 \end{bmatrix} + 8\begin{bmatrix} 2.08 & 0.06 \\ 0.06 & 0.17 \end{bmatrix} + 8\begin{bmatrix} 4.85 & -0.17 \\ -0.17 & 0.10 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 2.75 & 0.20 \\ 0.20 & 0.61 \end{bmatrix}, \tag{9.72a}$$

and using Eq. (9.66) the between-groups variance matrix is

$$[\mathbf{S}_{\text{B}}] = \frac{1}{2}\left(\begin{bmatrix} 12.96 & 5.33 \\ 5.33 & 2.19 \end{bmatrix} + \begin{bmatrix} 2.89 & -1.05 \\ -1.05 & 0.38 \end{bmatrix} + \begin{bmatrix} 32.49 & 5.81 \\ 5.81 & 1.04 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 24.17 & 5.04 \\ 5.04 & 1.81 \end{bmatrix}. \tag{9.72b}$$

The directions of the two discriminant functions are specified by the eigenvectors of the matrix

$$[\mathbf{D}] = [\mathbf{S}_{\text{pooled}}]^{-1}[\mathbf{S}_{\text{B}}]$$

$$= \begin{bmatrix} 0.373 & -0.122 \\ -0.122 & 1.685 \end{bmatrix}\begin{bmatrix} 24.17 & 5.04 \\ 5.04 & 1.81 \end{bmatrix} = \begin{bmatrix} 8.40 & 1.65 \\ 5.54 & 2.43 \end{bmatrix}, \tag{9.73a}$$

which, when scaled according to Eq. (9.67), are

$$\mathbf{d}_1 = \begin{bmatrix} 0.542 \\ 0.415 \end{bmatrix} \quad \text{and} \quad \mathbf{d}_2 = \begin{bmatrix} -0.282 \\ 1.230 \end{bmatrix}. \tag{9.73b}$$

The eigenvectors $\mathbf{d}_1$ and $\mathbf{d}_2$ define the directions of the first discriminant function $\delta_1 = \mathbf{d}_1^{\mathrm{T}}\mathbf{x}$ and the second discriminant function $\delta_2 = \mathbf{d}_2^{\mathrm{T}}\mathbf{x}$. Figure 9.12 shows the data for all three groups in Table 9.6 plotted in the discriminant space defined by these two functions. Points for groups 1 and 2 are shown by circles and X's, as in Fig. 9.11, and points for group 3 are shown by +'s. The heavy symbols locate the respective vector means for the three groups. Note that the point clouds for groups 1 and 2 appear to be stretched and distorted relative to their arrangement in Fig. 9.11. This is because the matrix **[D]** is not symmetric, so that the two discriminant vectors in Eq. (9.73b) are not orthogonal.

Fig. 9.12 Geometry of multiple discriminant analysis applied to the three groups of data in Table 9.6. Group 1 stations are plotted as circles, group 2 stations are plotted as Xs, and group 3 stations are plotted as +'s. The three vector means are indicated by the corresponding heavy symbols. The two axes are the first and second discriminant functions, and the heavy dashed lines divide sections of this discriminant space allocated to each group. The data for Atlanta and Augusta are misclassified as belonging to group 2. The two stations Columbus and New Orleans, which are not part of the training data in Table 9.6, are shown as question marks, and allocated to groups 3 and 1, respectively.

The heavy dashed lines in Fig. 9.12 divide the portions of the discriminant space that are assigned to each of the three groups by the classification criterion in Eq. (9.71). Here the data for Atlanta and Augusta have both been misclassified as belonging to group 2 rather than group 1. For Atlanta, for example, the squared distance to the group 1 mean is $[.542(78.6 - 80.6) + .415(4.73 - 5.67)]^2 + [-.282(78.6 - 80.6) + 1.230(4.73 - 5.67)]^2 = 2.52$, and the squared distance to the group 2 mean is $[.542(78.6 - 78.7) + .415(4.73 - 3.57)]^2 + [-.282(78.6 - 78.7) + 1.230(4.73 - 3.57)]^2 = 2.31$. A line in this discriminant space could be drawn visually that would include these two stations in the group 1 region. That the discriminant analysis has not specified this line is probably a consequence of the assumption of equal covariance matrices not being well satisfied.

The data points for the two stations Columbus and New Orleans, which are not part of the training data in Table 9.6, are shown by the question marks in Fig. 9.12. The location in the discriminant space of the point for New Orleans is $\delta_1 = (.542)(82.1) + (.415)(6.73) = 47.3$ and $\delta_2 = (-.282)(82.1) + (1.230)(6.73) = -14.9$, which is within the region assigned to group 1. The coordinates in discriminant space for the Columbus data are $\delta_1 = (.542)(74.7) + (.415)(3.37) = 41.9$ and $\delta_2 = (-.282)(74.7) + (1.230)(3.37) = -16.9$, which is within the region assigned to group 3. ¤

9.6 Cluster Analysis

Cluster analysis deals with separating data into groups whose identities are not known in advance. In general, even the "correct" number of groups into which the data should be sorted is not known in advance. Rather, it is the degree of similarity and difference between individual observations **x** that are used to define the groups, and to assign group membership. Examples of use of cluster analysis in the climatological literature include grouping daily weather observations into synoptic types (Kalkstein *et al.*, 1987), defining weather regimes from upper-air flow patterns (Mo and Ghil, 1988; Molteni *et al.*, 1990), grouping members of forecast ensembles (Tracton and Kalnay, 1993), grouping regions of the tropical oceans on the basis of ship observations (Wolter, 1987), and defining climatic regions based on surface climate variables (DeGaetano and Shulman, 1990; Fovell and Fovell, 1993; Galliani and Filippini, 1985; Guttman, 1993).

Cluster analysis is primarily an exploratory data analysis tool, rather than an inferential tool. Given a sample of data vectors **x** contained in the rows of a $(n \times K)$ data matrix **[X]**, the procedure will define groups and assign group membership at varying levels of aggregation. Unlike discriminant analysis, the procedure does not contain rules for assigning membership to future observations **y**. However, a cluster analysis can bring out groupings in the data that might otherwise be overlooked, possibly leading to empirically useful stratification of the data, or helping to suggest physical bases for observed structure in the data.

Most commonly implemented cluster analysis procedures are *hierarchical*. That is, they build a hierarchy of sets of groups, each of which is constructed by merging pairs of previously defined groups. The procedure begins by considering that the n observations of $\mathbf{x}$ have no group structure or, equivalently, that the data set consists of n groups containing one observation each. The first step is to find the two groups (i.e., data vectors) that are closest in their K-dimensional space, and to combine them into a new group. There are now $n - 1$ groups, one of which has two members. On each subsequent step, the two groups that are closest are merged to form a larger group. This process continues until, on the final, $(n - 1)$st, step all n observations have been aggregated into a single group.

The n-group clustering at the beginning of this process and the one-group clustering at the end of this process are neither useful nor enlightening. Hopefully, however, a natural clustering of the data into a workable number of informative groups will emerge at some intermediate stage. That is, one hopes that the n data vectors cluster or "clump together" in their K-dimensional space into some number G, $1 < G < n$, groups that reflect similar data generating processes. The ideal result is a division of the data that minimizes differences between members of a given cluster, and maximizes differences between members of different clusters.

9.6.1 Distance Measures and Clustering Methods

Central to the idea of the clustering of data points is the idea of distance. Clusters should consist of points separated by small distances, relative to the distances between clusters. The most intuitive and commonly used distance measure in cluster analysis is the Euclidean distance [Eq. (9.6)] in the K-dimensional space of the data vectors. Thus, the distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\| = \left[\sum_{k=1}^{K} (x_{i,k} - x_{j,k})^2\right]^{1/2}. \tag{9.74}$$

Euclidean distance is by no means the only available choice for measuring distance between points or clusters. For example, the squared distance $d_{i,j}^2$ might also be used. The angle between pairs of vectors [Eq. (9.7)] is another possible choice, as are the many other alternatives presented in Romesburg (1984). Tracton and Kalnay (1993) have used the anomaly correlation [Eq. (7.39)] to group members of forecast ensembles, and the ordinary Pearson correlation is sometimes used as a clustering criterion as well. These latter two criteria are inverse distance measures, which should be maximized within groups, and minimized between groups.

While distances between pairs of points can be unambiguously defined, there are a number of alternatives for defining the distance between clusters comprised of groups of points. The criterion used to define cluster-to-cluster (intercluster) distances essentially defines the method of clustering. A few of the most common clustering methods are listed below. Most of these can be used with distance measures other than that in Eq. (9.74) as well.

- *Single-linkage*, or minimum-distance clustering. Here the distance between clusters G_1 and G_2 is the smallest Euclidean distance between one member of G_1 and one member of G_2. Formally,

$$d_{G_1,G_2} = \min_{i \in G_1, j \in G_2} [d_{i,j}]. \tag{9.75}$$

- *Complete-linkage*, or maximum-distance clustering groups data points on the basis of the largest distance between points in the two groups G_1 and G_2,

$$d_{G_1,G_2} = \max_{i \in G_1, j \in G_2} [d_{i,j}]. \tag{9.76}$$

- *Average-linkage* clustering defines cluster-to-cluster distance as the average Euclidean distance between all possible pairs of points in the two groups being compared. If G_1 contains n_1 points and G_2 contains n_2 points, this distance measure is

$$d_{G_1,G_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{i,j}. \tag{9.77}$$

- *Centroid clustering* compares distances between the centroids, or vector averages, of pairs of clusters. According to this measure, the distance between G_1 and G_2 is

$$d_{G_1,G_2} = \|\bar{\mathbf{x}}_{G_1} - \bar{\mathbf{x}}_{G_2}\|, \tag{9.78}$$

  where the vector means are as defined in Eq. (9.67a).

- *Ward's minimum variance method* merges that pair of clusters that will result in the minimum sum of squared distances between the points and the centroids of their respective groups, summed over the resulting groups. That is, for all possible ways of merging two of $G + 1$ groups to make G groups, that merger is made that minimizes

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \|\mathbf{x}_i - \bar{\mathbf{x}}_g\|^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{k=1}^{K} (x_{i,k} - \bar{x}_{g,k})^2, \tag{9.79}$$

  where the centroid, or group mean, for the newly merged group is recomputed using the data for both of the previously separate groups, before the squared distances are calculated.

Figure 9.13 illustrates single-linkage, complete-linkage, and centroid clustering for two hypothetical groups G_1 and G_2 in a $K = 2$-dimensional space. The open circles denote data points, of which there are $n_1 = 2$ in G_1 and $n_2 = 3$ in G_2. The centroids of the two groups are indicated by the solid circles. The single-linkage distance between G_1 and G_2 is the distance $\|\mathbf{x}_2 - \mathbf{x}_3\|$ between the closest pair of points in the two groups. The complete-linkage distance is that between the most distant pair, $\|\mathbf{x}_1 - \mathbf{x}_5\|$. The centroid distance is the Euclidean distance between

Fig. 9.13 Three measures of the distance in $K = 2$-dimensional space, between a cluster G_1 containing the two elements $\mathbf{x}_1$ and $\mathbf{x}_2$, and a cluster G_2 containing the elements $\mathbf{x}_3$, $\mathbf{x}_4$, and $\mathbf{x}_5$. The data points are indicated by open circles, and centroids of the two groups are indicated by the solid circles. According to the maximum-distance, or complete-linkage, criterion, the distance between the two groups is $\|\mathbf{x}_1 - \mathbf{x}_5\|$, or the greatest distance between all six possible pairs of points in the two groups. The minimum-distance, or single-linkage, criterion computes the distance between the groups as equal to the distance between the nearest pair of points, or $\|\mathbf{x}_2 - \mathbf{x}_3\|$. According to the centroid method, the distance between the two clusters is the distance between the sample means of their respective points.

the two vector means $\|\mathbf{x}_{G_1} - \mathbf{x}_{G_2}\|$. The average-linkage distance can also be visualized in Fig. 9.13, as the average of the six possible distances between individual members of G_1 and G_2: $(\|\mathbf{x}_1 - \mathbf{x}_5\| + \|\mathbf{x}_1 - \mathbf{x}_4\| + \|\mathbf{x}_1 - \mathbf{x}_3\| + \|\mathbf{x}_2 - \mathbf{x}_5\| + \|\mathbf{x}_2 - \mathbf{x}_4\| + \|\mathbf{x}_2 - \mathbf{x}_3\|)/6$.

The results of a cluster analysis can depend strongly on which definition is chosen for the distances between clusters. Single-linkage clustering is susceptible to "chaining," or the production of a few large clusters, which are formed by virtue of nearness of pairs of points at opposite edges of clusters to be merged. At the other extreme, complete-linkage clusters tend to be more numerous, as the criterion for merging clusters is more stringent. Average-distance clustering appears to be intermediate between these two extremes. Kalkstein *et al.* (1987) compared average-linkage, centroid clustering, and Ward's method for clustering surface weather observations. They report that centroid clustering exhibited chaining, Ward's method tended to produce groups of roughly equal size, and that average-linkage clustering seemed to be the most satisfactory.

The average-distance criterion appears to be the most commonly used approach to clustering, but none of the available methods are guaranteed to be best in any

given circumstance. One approach to choosing clusters is to repeat the analysis using more than one clustering criterion, and hope that agreement among different methods reflects real structure in the data rather than artifacts of particular clustering criteria. In any case, the results of a cluster analysis should not be accepted blindly. Rather, they should make sense in terms of other information available to the analyst.

9.6.2 The Dendrogram, or Tree Diagram

The progress and intermediate results of a cluster analysis are conventionally illustrated using the *dendrogram*, or "tree" diagram. Beginning with the "twigs" at the beginning of the analysis, when each of the n observations $\mathbf{x}$ constitutes its own cluster, one pair of "branches" is joined at each step as the closest two clusters are merged. The distances between these clusters before they are merged are also indicated in the diagram by the distance of the points of merger from the initial n-cluster stage of the "twigs."

Figure 9.14 illustrates a simple dendrogram, reflecting the clustering of the five points plotted as open circles in Fig. 9.13. The analysis begins at the bottom of

Fig. 9.14 A dendrogram, or tree diagram, for a clustering of the five points plotted as open circles in Fig. 9.13. The results of the four clustering steps are indicated as the original five lines are progressively joined from the bottom to the top of the diagram, with the distances between joined clusters indicated by the heights of the horizontal lines.

Fig. 9.14, when all five points constitute separate clusters. At the first stage, the closest two points, $\mathbf{x}_3$ and $\mathbf{x}_4$ are merged into a new cluster. The distance $d_{3,4} = \|\mathbf{x}_3 - \mathbf{x}_4\|$ is proportional to the distance between the horizontal bar joining these two points and the bottom of the figure. At the next stage, the points $\mathbf{x}_1$ and $\mathbf{x}_2$ are merged into a single cluster because the distance between them is smallest of the six distances between the four clusters that existed at the previous stage. The distance $d_{1,2} = \|\mathbf{x}_1 - \mathbf{x}_2\|$ is necessarily larger than the distance $d_{3,4} = \|\mathbf{x}_3 - \mathbf{x}_4\|$, since $\mathbf{x}_1$ and $\mathbf{x}_2$ were not chosen for merger on the first step, and the horizontal line indicating the distance between them is plotted higher on Fig. 9.14 than the distance between $\mathbf{x}_3$ and $\mathbf{x}_4$. The third step merges $\mathbf{x}_5$ and the pair $(\mathbf{x}_3, \mathbf{x}_4)$, to yield the two-group stage indicated by the dashed lines in Fig. 9.13.

9.6.3 How Many Clusters?

A cluster analysis will produce a different grouping of n observations at each of the $n - 1$ steps. On the first step each observation is in a separate group, and on the last step all the observations are in a single group. An important practical problem in cluster analysis is the choice of which intermediate stage will be chosen as the final solution. Thus, one needs to choose the level of aggregation in the tree diagram at which to stop further merging of clusters. Although the goal guiding this choice is to find that level of clustering that maximizes similarity within clusters and minimizes similarity between clusters, in practice the best number of clusters for a given problem is seldom obvious. Generally the stopping point will require a subjective choice.

One approach to the problem of choosing the best number of clusters is through summary statistics based on ideas drawn from hypothesis testing. Some of these objective stopping criteria that have been developed are discussed in Fovell and Fovell (1993), who also provide references to the broader literature on such methods.

A traditional subjective approach to determination of the stopping level is to inspect a plot of the distances between merged clusters as a function of the stage of the analysis. When similar clusters are being merged early in the process, these distances are small, and they increase relatively little from step to step. Late in the process there may be only a few clusters, separated by large distances. If a point can be discerned where the distances between merged clusters jumps markedly, the process can be stopped just before these distances become large.

Wolter (1987) suggests a Monte Carlo approach, where sets of random numbers simulating the real data are subjected to cluster analysis. The distributions of clustering distances for the random numbers are compared to the actual clustering distances for the data of interest. The idea here is that genuine clusters in the real data should be closer than clusters in the random data, and that the clustering algorithm should be stopped at the point where clustering distances are greater than for the analysis of the random data.

Example 9.10. A Cluster Analysis in $K = 2$ Dimensions

As is the case with discriminant analysis, the mechanics of cluster analysis are easiest to see when the data vectors have only $K = 2$ dimensions. Consider the data in Table 9.6, where these two dimensions are average July temperature and average July precipitation. These data were collected into three groups for use in the discriminant analysis worked out in Example 9.9. However, the point of a cluster analysis is to try to discern group structure within a data set, without prior knowledge or information about the nature of that structure. Therefore, for purposes of the cluster analysis, the data in Table 9.6 should be regarded as consisting of $n = 28$ observations of two-dimensional vectors $\mathbf{x}$, whose natural groupings we would like to discern.

Because the temperature and precipitation values have different physical units, it is probably wise to normalize the data (i.e., convert them to standardized anomalies) before subjecting them to a clustering algorithm. That is, all the temperatures are transformed by subtracting the overall mean of the 28 temperatures (77.0° F), and dividing by the overall standard deviation of the temperature observations (4.42° F). For the precipitation values the corresponding mean and standard deviation are 4.19 and 1.36 in. The reason for this pretreatment of the data is to avoid the same kind of problem that can occur when conducting a PCA (Section 9.3) using unlike data, where a variable with a much higher variance than the others will dominate the analysis even if the high variance is an artifact of the units of measurement. In a cluster analysis, the relative sizes of the units can have a large impact on the Euclidean distance between points. For example, if the precipitation had been reported in millimeters, there would be apparently more distance between points in the direction of the precipitation axis, and a clustering algorithm would focus on precipitation differences to define groups. If the precipitation were reported in meters, there would be essentially no distance between points in the direction of the precipitation axis, and a clustering algorithm would separate points almost entirely on the basis of the temperature data.

Figure 9.15 shows the results of clustering the data in Table 9.6, using the Euclidean distance measure in Eq. (9.74), and the complete-linkage clustering criterion in Eq. (9.76). On the left is a tree diagram for the process, with the individual stations listed at the bottom as the "leaves." There are 27 horizontal lines in this tree diagram, each of which represents a merging of the two clusters it connects. For example, at the first stage of the analysis the two closest points (Springfield and St. Louis) are merged into the same cluster. At the second stage Huntsville and Athens are merged, at the third stage Worcester and Binghamton are merged, and at the fourth stage Macon and Augusta are merged. At the fifth stage Concordia is merged with the cluster consisting of Springfield and St. Louis.

The heights of the horizontal lines indicating group mergers correspond to the distances between the merged clusters. Since the merger at each stage is between the two closest clusters, these distances become greater at later stages.

Fig. 9.15 Dendrogram (a) and the corresponding plot of the distance between merged clusters as a function of the stage of the cluster analysis (b) for the data in Table 9.6. Normalized data (standardized anomalies) have been clustered according to the complete-linkage criterion. The distances between merged groups appear to increase markedly at stage 22 or 23, indicating that the analysis should stop after 21 or 22 stages, which for these data would yield seven or six clusters, respectively. The six numbered clusters correspond to the grouping of the data shown in Fig. 9.16. The seven-cluster solution would split Topeka and Kansas City from the Alabama and Georgia stations in G_5. The five-cluster solution would merge G_3 and G_4.

Figure 9.15b shows the distance between merged clusters as a function of the stage in the analysis. Subjectively, these distances climb gradually until perhaps stage 22 or stage 23, where the distances between combined clusters begin to become noticeably larger. A reasonable interpretation of this change in slope is that the "true" clusters have been defined at this point in the analysis, and that the larger distances at later stages indicate mergers of unlike clusters that should be distinct. Note, however, that a single change in slope does not occur in every cluster analysis, so that the choice of where to stop group mergers may not always be so clear-cut. It is possible, for example, for there to be two or more relatively flat regions in the plot of distance versus stage, separated by segments of larger slope. In such a case the choice of where to stop the analysis is more ambiguous.

If Fig. 9.15b is interpreted as exhibiting its first major slope increase between stages 22 and 23, a plausible point at which to stop the analysis would be after stage 22. This stopping point would result in the definition of the six clusters

Fig. 9.16 Scatterplot of the data in Table 9.6 expressed as standardized anomalies, with dashed lines showing the six groups defined in the cluster analysis tree diagram in Fig. 9.15. The five-group clustering would merge the central U.S. stations in groups 3 and 4. The seven-group clustering would split the two central U.S. stations in group 5 from six southeastern U.S. stations.

labeled G_1–G_6 on the tree diagram in Fig. 9.15a. This level of clustering assigns the nine northeastern stations (+ symbols) into two groups, assigns seven of the nine central stations (× symbols) into two groups, allocates the central stations Topeka and Kansas City to group 5 with six of the southeastern stations (○ symbols), and assigns the remaining four southeastern stations to a separate cluster.

Figure 9.16 indicates these six groups in the $K = 2$-dimensional space of the data, by separating points in each cluster with dashed lines. If this solution seemed too highly aggregated on the basis of the prior knowledge and information available to the analyst, one could choose the seven-cluster solution produced after stage 21, separating the central U.S. cities Topeka and Kansas City from the six southeastern cities in group 5. If the six-cluster solution seemed too finely split, the five-cluster solution produced after stage 23 would merge the central U.S. stations in groups 3 and 4. ¤

Exercises

9.1. Calculate the matrix product **[B][E]**, using the values in Eqs. (9.31) and (9.32).

9.2. Derive the regression equation produced in Example 6.1, using matrix notation.

9.3. Calculate the angle between the first two eigenvectors in Table 9.2, section a.

9.4. Regard the joint behavior of the maximum temperature data for Ithaca and Canandaigua in Table A.1 as being bivariate normal. The eigenvectors of the variance–covariance matrix for these two variables are $\mathbf{e}_1^T = [.700, .714]$ and $\mathbf{e}_2^T = [-.714, .700]$, where the first element of each vector corresponds to the Ithaca temperature.
   (a) Find the corresponding eigenvalues.
   (b) Plot the 75% and 95% confidence ellipses of this distribution.

9.5. Using information and results from Exercise 9.4:
   (a) Calculate the values of the first principal component for January 1 and January 2.
   (b) Estimate the variance of all 31 values of the first principal component.
   (c) What proportion of the total variability of the maximum temperature data is represented by the first principal component?

9.6. A PCA of the data in Table A.3 yields the three eigenvectors $\mathbf{e}_1^T = [.593, .552, -.587]$, $\mathbf{e}_2^T = [.332, -.831, -.446]$, and $\mathbf{e}_3^T = [.734, -.069, .676]$, where the three elements in each vector pertain to the temperature, precipitation, and pressure data, respectively. The corresponding three eigenvalues are $\lambda_1 = 2.476$, $\lambda_2 = 0.356$, and $\lambda_3 = 0.169$.
   (a) Was this analysis done using the covariance matrix or the correlation matrix? How can you tell?
   (b) How many principal components should be retained according to Kaiser's rule, and according to the broken-stick model?
   (c) Discuss how to find the number of principal components to retain according to rule N. If the appropriate computing resources are available, implement your algorithm.

9.7. Using the information in Table 9.5 and the data in Table A.1, calculate the values of the canonical variables v_1 and w_1 for January 6 and January 7.

9.8. Conduct a discriminant analysis using the temperature and pressure data in Table A.3, to classify data points as coming from either El Niño or non–El Niño years.
   (a) What is the discriminant vector, scaled to have unit length?
   (b) Which, if any, of the El Niño years have been misclassified?

Appendix A Example Data Sets

In real applications of climatological data analysis one would hope to use much more data (e.g., all available January daily data, rather than data for just a single year), and would have a computer perform the computations. These small data sets have been used in a number of examples in this book so that the computations can be performed by hand, and a clearer understanding of procedures can be achieved.

Table A.1

Daily Precipitation (inches) and Temperature[a] (°F) Observations at Ithaca and Canadaigua, New York, for January 1987

	Ithaca			Canandaigua		
Date	Precipitation	T_{max}	T_{min}	Precipitation	T_{max}	T_{min}
1	0.00	33	19	0.00	34	28
2	0.07	32	25	0.04	36	28
3	1.11	30	22	0.84	30	26
4	0.00	29	−1	0.00	29	19
5	0.00	25	4	0.00	30	16
6	0.00	30	14	0.00	35	24
7	0.00	37	21	0.02	44	26
8	0.04	37	22	0.05	38	24
9	0.02	29	23	0.01	31	24
10	0.05	30	27	0.09	33	29
11	0.34	36	29	0.18	39	29
12	0.06	32	25	0.04	33	27
13	0.18	33	29	0.04	34	31
14	0.02	34	15	0.00	39	26
15	0.02	53	29	0.06	51	38
16	0.00	45	24	0.03	44	23
17	0.00	25	0	0.04	25	13
18	0.00	28	2	0.00	34	14

continues

Table A.1

Continued

	Ithaca			Canandaigua		
Date	Precipitation	T_{max}	T_{min}	Precipitation	T_{max}	T_{min}
19	0.00	32	26	0.00	36	28
20	0.45	27	17	0.35	29	19
21	0.00	26	19	0.02	27	19
22	0.00	28	9	0.01	29	17
23	0.70	24	20	0.35	27	22
24	0.00	26	−6	0.08	24	2
25	0.00	9	−13	0.00	11	4
26	0.00	22	−13	0.00	21	5
27	0.00	17	−11	0.00	19	7
28	0.00	26	−4	0.00	26	8
29	0.01	27	−4	0.01	28	14
30	0.03	30	11	0.01	31	14
31	0.05	34	23	0.13	38	23
Sum or average	3.15	29.87	13.00	2.40	31.77	20.23

[a] T_{max}—maximum temperature; T_{min}—minimum temperature.

Table A.2

January Precipitation at Ithaca, New York, 1993–1982, inches

1933	0.44	1945	2.74	1958	4.90	1970	1.03
1934	1.18	1946	1.13	1959	2.94	1971	1.11
1935	2.69	1947	2.50	1960	1.75	1972	1.35
1936	2.08	1948	1.72	1961	1.69	1973	1.44
1937	3.66	1949	2.27	1962	1.88	1974	1.84
1938	1.72	1950	2.82	1963	1.31	1975	1.69
1939	2.82	1951	1.98	1964	1.76	1976	3.00
1940	0.72	1952	2.44	1965	2.17	1977	1.36
1941	1.46	1953	2.53	1966	2.38	1978	6.37
1942	1.30	1954	2.00	1967	1.16	1979	4.55
1943	1.35	1955	1.12	1968	1.39	1980	0.52
1944	0.54	1956	2.13	1969	1.36	1981	0.87
		1957	1.36			1982	1.51

Table A.3

June Climatic Data for Guayaquil, Ecuador, 1951–1970
(Asterisks Indicate El Niño Years)

Year	Temperature, °C	Precipitation, mm	Pressure, mb
1951*	26.1	43	1009.5
1952	24.5	10	1010.9
1953*	24.8	4	1010.7
1954	24.5	0	1011.2
1955	24.1	2	1011.9
1956	24.3	Missing	1011.2
1957*	26.4	31	1009.3
1958	24.9	0	1011.1
1959	23.7	0	1012.0
1960	23.5	0	1011.4
1961	24.0	2	1010.9
1962	24.1	3	1011.5
1963	23.7	0	1011.0
1964	24.3	4	1011.2
1965*	26.6	15	1009.9
1966	24.6	2	1012.5
1967	24.8	0	1011.1
1968	24.4	1	1011.8
1969*	26.8	127	1009.3
1970	25.2	2	1010.6

Appendix B | Probability Tables

The following tables list selected common probability distributions for which closed-form expressions for the cumulative distribution functions do not exist.

Table B.1

Left-Tail Cumulative Probabilities for the Standard Gaussian Distribution, $\Phi(z) = \Pr\{Z \leq z\}$[a]

Z	.09	.08	.07	.06	.05	.04	.03	.02	.01	.00	Z
−4.0	.00002	.00002	.00002	.00002	.00003	.00003	.00003	.00003	.00003	.00003	−4.0
−3.9	.00003	.00003	.00004	.00004	.00004	.00004	.00004	.00004	.00005	.00005	−3.9
−3.8	.00005	.00005	.00005	.00006	.00006	.00006	.00006	.00007	.00007	.00007	−3.8
−3.7	.00008	.00008	.00008	.00008	.00009	.00009	.00010	.00010	.00010	.00011	−3.7
−3.6	.00011	.00012	.00012	.00013	.00013	.00014	.00014	.00015	.00015	.00016	−3.6
−3.5	.00017	.00017	.00018	.00019	.00019	.00020	.00021	.00022	.00022	.00023	−3.5
−3.4	.00024	.00025	.00026	.00027	.00028	.00029	.00030	.00031	.00032	.00034	−3.4
−3.3	.00035	.00036	.00038	.00039	.00040	.00042	.00043	.00045	.00047	.00048	−3.3
−3.2	.00050	.00052	.00054	.00056	.00058	.00060	.00062	.00064	.00066	.00069	−3.2
−3.1	.00071	.00074	.00076	.00079	.00082	.00084	.00087	.00090	.00094	.00097	−3.1
−3.0	.00100	.00104	.00107	.00111	.00114	.00118	.00122	.00126	.00131	.00135	−3.0
−2.9	.00139	.00144	.00149	.00154	.00159	.00164	.00169	.00175	.00181	.00187	−2.9
−2.8	.00193	.00199	.00205	.00212	.00219	.00226	.00233	.00240	.00248	.00256	−2.8
−2.7	.00264	.00272	.00280	.00289	.00298	.00307	.00317	.00326	.00336	.00347	−2.7
−2.6	.00357	.00368	.00379	.00391	.00402	.00415	.00427	.00440	.00453	.00466	−2.6
−2.5	.00480	.00494	.00508	.00523	.00539	.00554	.00570	.00587	.00604	.00621	−2.5
−2.4	.00639	.00657	.00676	.00695	.00714	.00734	.00755	.00776	.00798	.00820	−2.4
−2.3	.00842	.00866	.00889	.00914	.00939	.00964	.00990	.01017	.01044	.01072	−2.3
−2.2	.01101	.01130	.01160	.01191	.01222	.01255	.01287	.01321	.01355	.01390	−2.2
−2.1	.01426	.01463	.01500	.01539	.01578	.01618	.01659	.01700	.01743	.01786	−2.1
−2.0	.01831	.01876	.01923	.01970	.02018	.02068	.02118	.02169	.02222	.02275	−2.0
−1.9	.02330	.02385	.02442	.02500	.02559	.02619	.02680	.02743	.02807	.02872	−1.9
−1.8	.02938	.03005	.03074	.03144	.03216	.03288	.03362	.03438	.03515	.03593	−1.8

continues

Table B.1

Continued

Z	.09	.08	.07	.06	.05	.04	.03	.02	.01	.00	Z
−1.7	.03673	.03754	.03836	.03920	.04006	.04093	.04182	.04272	.04363	.04457	−1.7
−1.6	.04551	.04648	.04746	.04846	.04947	.05050	.05155	.05262	.05370	.05480	−1.6
−1.5	.05592	.05705	.05821	.05938	.06057	.06178	.06301	.06426	.06552	.06681	−1.5
−1.4	.06811	.06944	.07078	.07215	.07353	.07493	.07636	.07780	.07927	.08076	−1.4
−1.3	.08226	.08379	.08534	.08692	.08851	.09012	.09176	.09342	.09510	.09680	−1.3
−1.2	.09853	.10027	.10204	.10383	.10565	.10749	.10935	.11123	.11314	.11507	−1.2
−1.1	.11702	.11900	.12100	.12302	.12507	.12714	.12924	.13136	.13350	.13567	−1.1
−1.0	.13786	.14007	.14231	.14457	.14686	.14917	.15151	.15386	.15625	.15866	−1.0
−0.9	.16109	.16354	.16602	.16853	.17106	.17361	.17619	.17879	.18141	.18406	−0.9
−0.8	.18673	.18943	.19215	.19489	.19766	.20045	.20327	.20611	.20897	.21186	−0.8
−0.7	.21476	.21770	.22065	.22363	.22663	.22965	.23270	.23576	.23885	.24196	−0.7
−0.6	.24510	.24825	.25143	.25463	.25785	.26109	.26435	.26763	.27093	.27425	−0.6
−0.5	.27760	.28096	.28434	.28774	.29116	.29460	.29806	.30153	.30503	.30854	−0.5
−0.4	.31207	.31561	.31918	.32276	.32636	.32997	.33360	.33724	.34090	.34458	−0.4
−0.3	.34827	.35197	.35569	.35942	.36317	.36693	.37070	.37448	.37828	.38209	−0.3
−0.2	.38591	.38974	.39358	.39743	.40129	.40517	.40905	.41294	.41683	.42074	−0.2
−0.1	.42465	.42858	.43251	.43644	.44038	.44433	.44828	.45224	.45620	.46017	−0.1
−0.0	.46414	.46812	.47210	.47608	.48006	.48405	.48803	.49202	.49601	.50000	0.0

[a] Values of the standardized Gaussian variable, z, are listed to tenths in the rightmost and leftmost columns. Remaining column headings index the hundredth place of z. Right-tail probabilities are obtained using $\Pr\{Z > z\} = 1 - \Pr\{Z \leq z\}$. Probabilities for $Z > 0$ are obtained using the symmetry of the Gaussian distribution, $\Pr\{Z \leq z\} = 1 - \Pr\{Z \leq -z\}$.

Table B.2

Quantiles of the Standard ($\beta = 1$) Gamma Distribution[a]

	Cumulative probability														
α	.001	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	.999
0.05	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.077	0.262	1.057	2.423
0.10	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000	0.001	0.004	0.018	0.070	0.264	0.575	1.554	3.035
0.15	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.001	0.006	0.021	0.062	0.164	0.442	0.820	1.894	3.439
0.20	0.0000	0.0000	0.0000	0.000	0.000	0.002	0.007	0.021	0.053	0.122	0.265	0.602	1.024	2.164	3.756
0.25	0.0000	0.0000	0.0000	0.000	0.001	0.006	0.018	0.044	0.095	0.188	0.364	0.747	1.203	2.395	4.024
0.30	0.0000	0.0000	0.0000	0.000	0.003	0.013	0.034	0.073	0.142	0.257	0.461	0.882	1.365	2.599	4.262
0.35	0.0000	0.0000	0.0001	0.001	0.007	0.024	0.055	0.108	0.192	0.328	0.556	1.007	1.515	2.785	4.477
0.40	0.0000	0.0000	0.0004	0.002	0.013	0.038	0.080	0.145	0.245	0.398	0.644	1.126	1.654	2.958	4.677
0.45	0.0000	0.0000	0.0010	0.005	0.022	0.055	0.107	0.186	0.300	0.468	0.733	1.240	1.786	3.121	4.863
0.50	0.0000	0.0001	0.0020	0.008	0.032	0.074	0.138	0.228	0.355	0.538	0.819	1.349	1.913	3.274	5.040
0.55	0.0000	0.0002	0.0035	0.012	0.045	0.096	0.170	0.272	0.411	0.607	0.904	1.454	2.034	3.421	5.208
0.60	0.0000	0.0004	0.0057	0.018	0.059	0.120	0.204	0.316	0.467	0.676	0.987	1.556	2.150	3.562	5.370
0.65	0.0000	0.0008	0.0086	0.025	0.075	0.146	0.240	0.362	0.523	0.744	1.068	1.656	2.264	3.698	5.526
0.70	0.0001	0.0013	0.0123	0.033	0.093	0.173	0.276	0.408	0.579	0.811	1.149	1.753	2.374	3.830	5.676
0.75	0.0001	0.0020	0.0168	0.043	0.112	0.201	0.314	0.455	0.636	0.878	1.227	1.848	2.481	3.958	5.822
0.80	0.0003	0.0030	0.0221	0.053	0.132	0.231	0.352	0.502	0.692	0.945	1.305	1.941	2.586	4.083	5.964
0.85	0.0004	0.0044	0.0283	0.065	0.153	0.261	0.391	0.550	0.749	1.010	1.382	2.032	2.689	4.205	6.103
0.90	0.0007	0.0060	0.0353	0.078	0.176	0.292	0.431	0.598	0.805	1.076	1.458	2.122	2.790	4.325	6.239
0.95	0.0010	0.0080	0.0432	0.091	0.199	0.324	0.471	0.646	0.861	1.141	1.533	2.211	2.888	4.441	6.373
1.00	0.0014	0.0105	0.0517	0.106	0.224	0.357	0.512	0.694	0.918	1.206	1.607	2.298	2.986	4.556	6.503
1.05	0.0019	0.0133	0.0612	0.121	0.249	0.391	0.553	0.742	0.974	1.270	1.681	2.384	3.082	4.669	6.631
1.10	0.0022	0.0166	0.0713	0.138	0.275	0.425	0.594	0.791	1.030	1.334	1.759	2.469	3.177	4.781	6.757
1.15	0.0023	0.0202	0.0823	0.155	0.301	0.459	0.636	0.840	1.086	1.397	1.831	2.553	3.270	4.890	6.881
1.20	0.0024	0.0240	0.0938	0.173	0.329	0.494	0.678	0.889	1.141	1.460	1.903	2.636	3.362	4.998	7.003

continues

Table B.2

	Cumulative probability														
α	.001	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	.999
1.25	0.0031	0.0271	0.1062	0.191	0.357	0.530	0.720	0.938	1.197	1.523	1.974	2.719	3.453	5.105	7.124
1.30	0.0037	0.0321	0.1192	0.210	0.385	0.566	0.763	0.987	1.253	1.586	2.045	2.800	3.544	5.211	7.242
1.35	0.0044	0.0371	0.1328	0.230	0.414	0.602	0.806	1.036	1.308	1.649	2.115	2.881	3.633	5.314	7.360
1.40	0.0054	0.0432	0.1451	0.250	0.443	0.639	0.849	1.085	1.364	1.711	2.185	2.961	3.722	5.418	7.476
1.45	0.0066	0.0493	0.1598	0.272	0.473	0.676	0.892	1.135	1.419	1.773	2.255	3.041	3.809	5.519	7.590
1.50	0.0083	0.0560	0.1747	0.293	0.504	0.713	0.935	1.184	1.474	1.834	2.324	3.120	3.897	5.620	7.704
1.55	0.0106	0.0632	0.1908	0.313	0.534	0.750	0.979	1.234	1.530	1.896	2.392	3.199	3.983	5.720	7.816
1.60	0.0136	0.0708	0.2070	0.336	0.565	0.788	1.023	1.283	1.585	1.957	2.461	3.276	4.068	5.818	7.928
1.65	0.0177	0.0780	0.2238	0.359	0.597	0.826	1.067	1.333	1.640	2.018	2.529	3.354	4.153	5.917	8.038
1.70	0.0232	0.0867	0.2411	0.382	0.628	0.865	1.111	1.382	1.695	2.079	2.597	3.431	4.237	6.014	8.147
1.75	0.0306	0.0958	0.2588	0.406	0.661	0.903	1.155	1.432	1.750	2.140	2.664	3.507	4.321	6.110	8.255
1.80	0.0360	0.1041	0.2771	0.430	0.693	0.942	1.199	1.481	1.805	2.200	2.731	3.584	4.405	6.207	8.362
1.85	0.0406	0.1145	0.2958	0.454	0.726	0.980	1.244	1.531	1.860	2.261	2.798	3.659	4.487	6.301	8.469
1.90	0.0447	0.1243	0.3142	0.479	0.759	1.020	1.288	1.580	1.915	2.321	2.865	3.735	4.569	6.396	8.575
1.95	0.0486	0.1361	0.3338	0.505	0.790	1.059	1.333	1.630	1.969	2.381	2.931	3.809	4.651	6.490	8.679
2.00	0.0525	0.1514	0.3537	0.530	0.823	1.099	1.378	1.680	2.024	2.442	2.997	3.883	4.732	6.582	8.783
2.05	0.0565	0.1637	0.3741	0.556	0.857	1.138	1.422	1.729	2.079	2.501	3.063	3.958	4.813	6.675	8.887
2.10	0.0657	0.1751	0.3949	0.583	0.891	1.178	1.467	1.779	2.133	2.561	3.129	4.032	4.894	6.767	8.989
2.15	0.0697	0.1864	0.4149	0.610	0.925	1.218	1.512	1.829	2.188	2.620	3.195	4.105	4.973	6.858	9.091
2.20	0.0740	0.2002	0.4365	0.637	0.959	1.258	1.557	1.879	2.242	2.680	3.260	4.179	5.053	6.949	9.193
2.25	0.0854	0.2116	0.4584	0.664	0.994	1.298	1.603	1.928	2.297	2.739	3.325	4.252	5.132	7.039	9.294
2.30	0.0898	0.2259	0.4807	0.691	1.029	1.338	1.648	1.978	2.351	2.799	3.390	4.324	5.211	7.129	9.394
2.35	0.0945	0.2378	0.5023	0.718	1.064	1.379	1.693	2.028	2.405	2.858	3.455	4.396	5.289	7.219	9.493
2.40	0.0996	0.2526	0.5244	0.747	1.099	1.420	1.738	2.078	2.459	2.917	3.519	4.468	5.367	7.308	9.592
2.45	0.1134	0.2680	0.5481	0.775	1.134	1.460	1.784	2.127	2.514	2.976	3.584	4.540	5.445	7.397	9.691
2.50	0.1184	0.2803	0.5754	0.804	1.170	1.500	1.829	2.178	2.568	3.035	3.648	4.612	5.522	7.484	9.789

2.55	0.1239	0.2962	0.5978	0.833	1.205	1.539	1.875	2.227	2.622	3.093	3.712	4.683	5.600	7.572	9.886
2.60	0.1297	0.3129	0.6211	0.862	1.241	1.581	1.920	2.277	2.676	3.152	3.776	4.754	5.677	7.660	9.983
2.65	0.1468	0.3255	0.6456	0.890	1.277	1.622	1.966	2.327	2.730	3.210	3.840	4.825	5.753	7.746	10.079
2.70	0.1523	0.3426	0.6705	0.920	1.314	1.663	2.011	2.376	2.784	3.269	3.903	4.896	5.830	7.833	10.176
2.75	0.1583	0.3561	0.6938	0.950	1.350	1.704	2.058	2.427	2.838	3.328	3.967	4.966	5.906	7.919	10.272
2.80	0.1647	0.3735	0.7188	0.980	1.386	1.746	2.103	2.476	2.892	3.386	4.030	5.040	5.982	8.004	10.367
2.85	0.1861	0.3919	0.7441	1.009	1.423	1.787	2.149	2.526	2.946	3.444	4.093	5.120	6.058	8.090	10.461
2.90	0.1919	0.4056	0.7697	1.040	1.460	1.829	2.195	2.576	2.999	3.502	4.156	5.190	6.133	8.175	10.556
2.95	0.1982	0.4242	0.7936	1.070	1.497	1.871	2.241	2.626	3.054	3.560	4.220	5.260	6.208	8.260	10.649
3.00	0.2050	0.4388	0.8193	1.101	1.534	1.913	2.287	2.676	3.108	3.618	4.283	5.329	6.283	8.345	10.743
3.05	0.2123	0.4577	0.8454	1.134	1.571	1.954	2.333	2.726	3.161	3.676	4.346	5.398	6.357	8.429	10.837
3.10	0.2385	0.4778	0.8717	1.165	1.607	1.996	2.378	2.776	3.215	3.734	4.408	5.468	6.432	8.513	10.930
3.15	0.2447	0.4922	0.8982	1.197	1.645	2.038	2.425	2.825	3.268	3.792	4.471	5.537	6.506	8.596	11.023
3.20	0.2514	0.5125	0.9251	1.227	1.682	2.080	2.471	2.875	3.322	3.850	4.533	5.605	6.580	8.680	11.113
3.25	0.2588	0.5278	0.9498	1.259	1.720	2.123	2.517	2.925	3.376	3.907	4.595	5.675	6.654	8.763	11.205
3.30	0.2667	0.5483	0.9767	1.291	1.758	2.165	2.563	2.975	3.430	3.965	4.658	5.743	6.727	8.845	11.298
3.35	0.2995	0.5704	1.0039	1.323	1.796	2.207	2.610	3.025	3.483	4.022	4.720	5.811	6.801	8.928	11.389
3.40	0.3057	0.5850	1.0313	1.354	1.834	2.250	2.656	3.075	3.537	4.079	4.782	5.879	6.874	9.010	11.480
3.45	0.3126	0.6072	1.0590	1.386	1.872	2.292	2.702	3.125	3.590	4.137	4.843	5.948	6.947	9.093	11.570
3.50	0.3201	0.6228	1.0870	1.418	1.910	2.334	2.748	3.175	3.644	4.194	4.905	6.015	7.020	9.174	11.660
3.55	0.3282	0.6450	1.1152	1.451	1.948	2.377	2.795	3.225	3.697	4.252	4.967	6.084	7.092	9.255	11.749
3.60	0.3370	0.6614	1.1405	1.483	1.985	2.420	2.841	3.274	3.750	4.309	5.028	6.152	7.165	9.337	11.840
3.65	0.3767	0.6837	1.1687	1.516	2.024	2.462	2.887	3.324	3.804	4.366	5.091	6.219	7.237	9.418	11.929
3.70	0.3830	0.7084	1.1972	1.549	2.062	2.505	2.934	3.374	3.858	4.423	5.152	6.286	7.310	9.499	12.017
3.75	0.3900	0.7233	1.2259	1.582	2.101	2.547	2.980	3.425	3.911	4.480	5.214	6.354	7.381	9.579	12.107
3.80	0.3978	0.7480	1.2549	1.613	2.140	2.590	3.027	3.474	3.964	4.537	5.275	6.420	7.454	9.659	12.195
3.85	0.4064	0.7637	1.2843	1.646	2.179	2.633	3.073	3.524	4.018	4.594	5.336	6.488	7.525	9.740	12.284
3.90	0.4157	0.7883	1.3101	1.680	2.218	2.676	3.120	3.574	4.071	4.651	5.397	6.555	7.596	9.820	12.371
3.95	0.4259	0.8049	1.3393	1.713	2.257	2.719	3.163	3.624	4.124	4.708	5.458	6.622	7.668	9.900	12.459
4.00	0.4712	0.8294	1.3687	1.746	2.295	2.762	3.209	3.674	4.177	4.765	5.519	6.689	7.739	9.980	12.546

continues

Table B.2

Continued

	Cumulative probability														
α	.001	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99	.999
4.05	0.4779	0.8469	1.3984	1.780	2.334	2.805	3.256	3.724	4.231	4.822	5.580	6.755	7.811	10.059	12.634
4.10	0.4853	0.8714	1.4285	1.814	2.373	2.848	3.302	3.774	4.284	4.879	5.641	6.821	7.882	10.137	12.721
4.15	0.4937	0.8999	1.4551	1.848	2.413	2.891	3.350	3.823	4.337	4.936	5.701	6.888	7.952	10.216	12.807
4.20	0.5030	0.9141	1.4850	1.882	2.451	2.935	3.396	3.874	4.390	4.992	5.762	6.954	8.023	10.295	12.894
4.25	0.5133	0.9424	1.5150	1.916	2.491	2.978	3.443	3.924	4.444	5.049	5.823	7.020	8.093	10.374	12.981
4.30	0.5244	0.9575	1.5454	1.950	2.531	3.021	3.489	3.974	4.497	5.105	5.883	7.086	8.170	10.453	13.066
4.35	0.5779	0.9856	1.5762	1.985	2.572	3.065	3.537	4.024	4.550	5.162	5.944	7.153	8.264	10.531	13.152
4.40	0.5842	1.0016	1.6034	2.017	2.612	3.108	3.584	4.074	4.603	5.218	6.005	7.219	8.334	10.609	13.238
4.45	0.5916	1.0294	1.6339	2.051	2.653	3.152	3.630	4.123	4.656	5.274	6.065	7.284	8.405	10.687	13.324
4.50	0.6001	1.0463	1.6646	2.085	2.691	3.195	3.677	4.173	4.709	5.331	6.126	7.350	8.475	10.765	13.410
4.55	0.6096	1.0739	1.6956	2.120	2.731	3.239	3.724	4.223	4.762	5.387	6.186	7.415	8.544	10.843	13.495
4.60	0.6202	1.0917	1.7271	2.155	2.771	3.283	3.771	4.273	4.815	5.443	6.246	7.480	8.615	10.920	13.578
4.65	0.6319	1.1191	1.7547	2.190	2.812	3.326	3.817	4.323	4.868	5.501	6.306	7.546	8.684	10.998	13.663
4.70	0.6978	1.1378	1.7857	2.225	2.852	3.369	3.864	4.373	4.921	5.557	6.366	7.611	8.754	11.075	13.748
4.75	0.7031	1.1649	1.8170	2.260	2.890	3.412	3.911	4.423	4.974	5.613	6.426	7.676	8.823	11.152	13.832
4.80	0.7095	1.1844	1.8487	2.295	2.930	3.456	3.958	4.474	5.027	5.669	6.486	7.742	8.892	11.229	13.916
4.85	0.7172	1.2113	1.8809	2.330	2.970	3.500	4.005	4.524	5.081	5.725	6.546	7.807	8.962	11.306	14.000
4.90	0.7262	1.2465	1.9088	2.366	3.011	3.544	4.052	4.573	5.134	5.781	6.606	7.872	9.031	11.382	14.084
4.95	0.7365	1.2582	1.9403	2.398	3.051	3.588	4.099	4.623	5.186	5.837	6.665	7.937	9.100	11.457	14.168
5.00	0.7482	1.2931	1.9722	2.434	3.091	3.632	4.146	4.673	5.239	5.893	6.725	8.002	9.169	11.534	14.251

[a] Tabulated elements are values of the standardized random variable ξ corresponding to the cumulative probabilities $F(\xi)$ given in the column headings, for values of the shape parameter (α) given in the first column. To find quantiles for distributions with other scale parameters, enter the table at the appropriate row, read the standardized value in the appropriate column, and multiply the tabulated value by the scale parameter. To extract cumulative probabilities corresponding to a given value of the random variable, divide the value by the scale parameter, enter the table at the row appropriate to the shape parameter, and interpolate the result from the column headings. The χ^2 distribution is a special case of the gamma, with $\alpha = \nu/2$ and $\beta = 2$, so its probabilities can be found using the rows for $\alpha = 0.5, 1.0, 1.5$, and so on.

Appendix C | Answers to Exercises

Chapter 2

2.1. (b) $\Pr\{A \cup B\} = 0.7$
(c) $\Pr\{A \cap B^C\} = 0.1$
(d) $\Pr\{A^C \cap B^C\} = 0.3$

2.2. (b) $\Pr\{A\} = 9/31$, $\Pr\{B\} = 15/31$, $\Pr\{A,B\} = 9/31$
(c) $\Pr\{A \mid B\} = 9/15$
(d) No: $\Pr\{A\} \neq \Pr\{A \mid B\}$

2.3. (b) $\Pr\{E_1, E_2, E_3\} = 0.000125$
(c) $\Pr\{E_1^C, E_2^C, E_3^C\} = 0.857$

2.4. 0.20

Chapter 3

3.1. median = 2 mm, trimean = 2.75 mm, mean = 12.95 mm

3.2. MAD = 0.4 mb, IQR = 0.8 mb, $s = 0.88$ mb

3.4. $\gamma_{YK} = 0.158$, $\gamma = 0.877$

3.7. $\lambda = 0$

3.9. $z = 1.36$

3.10. $r_0 = 1.000$, $r_1 = 0.652$, $r_2 = 0.388$, $r_3 = 0.281$

3.12. Pearson:

$$\begin{matrix} 1.000 & & \\ 0.703 & 1.000 & \\ -0.830 & -0.678 & 1.000 \end{matrix}$$

Spearman:

$$\begin{matrix} 1.000 & & \\ 0.606 & 1.000 & \\ -0.688 & -0.632 & 1.000 \end{matrix}$$

Chapter 4

4.1. 0.185

4.2. (a) 0.041
(b) 0.292

4.3. (a) $\mu_{\text{drought}} = 0.056$, $\mu_{\text{wet}} = 0.565$
(b) 0.054
(c) 0.432

4.4. \$280 million, \$2.825 billion
4.5. (a) $\mu = 24.8°C$, $\sigma = 0.98°C$
(b) $\mu = 76.6°F$, $\sigma = 1.76°F$
4.6. (a) 0.00939
(b) 22.9°C
4.7. (a) $\alpha = 3.785$, $\beta = 0.934$ in.
(b) $\alpha = 3.785$, $\beta = 23.7$ mm
4.8. (a) $q_{30} = 2.41$ in. = 61.2 mm; $q_{70} = 4.22$ in. = 107.22 mm
(b) 0.30 in., or 7.7 mm
(c) $\cong 0.05$
4.9. (a) $q_{30} = 2.30$ in. = 58.3 mm; $q_{70} = 4.13$ in. = 104.9 mm
(b) 0.46 in., or 11.6 mm
(c) $\cong 0.07$
4.10. (a) $\beta = 35.1$ cm, $\xi = 59.7$ cm
(b) $x = \xi - \beta \ln[-\ln(F)]$; $\Pr\{X \leq 221 \text{ cm}\} = 0.99$
4.11. (a) $\mu_{\max} = 31.8°F$, $\sigma_{\max} = 7.86°F$, $\mu_{\min} = 20.2°F$, $\sigma_{\min} = 8.81°F$, $\rho = 0.810$
(b) 0.728

Chapter 5

5.1. (a) $z = 6.33$, reject H_0
(b) [1.26°C, 2.40°C]
5.2. 6.53 days (Ithaca), 6.08 days (Canandaigua)
5.3. $z = 2.95$
(a) $p = 0.00318$
(b) $p = 0.00159$
5.4. $|r| \geq 0.366$
5.5. (a) $D_n = 0.152$ (not reject at 5% level)
(b) For classes: [<2, 2–3, 3–4, 4–5, ≥5], $\chi^2 = 0.327$ (do not reject)
5.6. $\Lambda = 21.86$, reject ($p < .001$)
5.7. (a) $U_1 = 1$, reject ($p < .005$)
(b) $z = -3.18$, reject ($p = .00074$)
5.8. $\approx [1.02, 3.59]$
5.9. Observed $(s^2_{\text{E-N}}/s^2_{\text{non-E-N}}) = 329.5$, critical value (1%, two-tailed) ≈ 141, reject ($p < .01$)
5.10. Not significant, even assuming spatial independence ($p = 0.19$, one-tailed)

Chapter 6

6.1. (a) $a = 959.8°C$, $b = -0.925°C/mb$
(b) $z = -6.33$

(c) 0.690
(d) 0.881
(e) 0.925

6.2. (a) 3
(b) 117.9
(c) 0.974
(d) 0.715

6.3. (a) 1.74 mm
(b) [0 mm, 14 mm]

6.4. Range of slopes, −0.850 to −1.095; MSE = 0.369

6.5. (a) −59 nautical miles
(b) −66 nautical miles

6.6. (a) 65.8°F
(b) 52.5°F
(c) 21.7°F
(d) 44.5°F

6.7. (a) 0.65
(b) 0.49
(c) 0.72
(d) 0.56

6.8. $f_{MOS} = 30.8°F + (0)(Th)$

6.9. 0.20

6.10. (a) 12 mm
(b) [5 mm, 32 mm], [1 mm, 55 mm]
(c) 0.625

Chapter 7

7.1. (a)

.0025	.0013	.0108	.0148	.0171	.0138
.0155	.0161	.0177	.0176	.0159	.0189
.4087	.0658	.1725	.0838	.0445	.0228
.0148	.0114	.0068	.0044	.0011	.0014

(b) 0.162

7.2. 1644 1330
364 9064

7.3. (a) 0.863
(b) 0.493
(c) 0.578
(d) 0.691

7.4. (a) 0.074
(b) 0.097
(c) 0.761
(d) 0.406

7.5. (a)

.9597	.0127	.0021	.0007
.0075	.0043	.0014	.0005
.0013	.0013	.0009	.0003
.0007	.0006	.0049	.0024

(b) 0.966
(c) 0.369
(d) 0.334

7.6. (a) 5.37°F
(b) 7.54°F
(c) −0.03°F
(d) 0.66%

7.7. (a) 0.1215
(b) 0.1699
(c) 28.5%

7.8. (a) 0.298
(b) 16.4%

7.9. (a) 30.3
(b) 5.31 dam^2
(c) 46.9%
(d) 0.714

Chapter 8

8.1. (a) $p_{01} = 0.45, p_{11} = 0.79$
(b) $\chi^2 = 3.51, p \approx 0.064$
(c) $\pi_1 = 0.682, n_{\cdot 1}/n = 0.667$
(d) $r_0 = 1.00, r_1 = 0.34, r_2 = 0.12, r_3 = 0.04$

8.2. (a) $r_0 = 1.00, r_1 = 0.40, r_2 = 0.16, r_3 = 0.06, r_4 = 0.03, r_5 = 0.01$
(b) $r_0 = 1.00, r_1 = 0.41, r_2 = -0.41, r_3 = -0.58, r_4 = -0.12, r_5 = 0.32$

8.3. (a) AR(1): $\phi = 0.80; s_\varepsilon^2 = 64.0$
AR(2): $\phi_1 = 0.89, \phi_2 = -0.11; s_\varepsilon^2 = 35.5$
AR(3): $\phi_1 = 0.91, \phi_2 = -0.25, \phi_3 = 0.16; s_\varepsilon^2 = 34.7$
(b) AR(2): BIC = 373.8
(c) AR(2): AIC = 366.0

8.4. $x_1 = 71.5, x_2 = 66.3, x_3 = 62.1$

8.5. (a) 28.6
(b) 19.8
(c) 4.5

8.6. (a) $C_1 = 16.92$°F, $\phi_1 = 199°$; $C_2 = 4.16$°F, $\phi_2 = 256°$

8.7. (a) 82.0°F
(b) 74.8°F

8.8. (b) 0.990

8.9. 56

8.10. (a) For example, $f_A = 1 - .0508 \text{ month}^{-1} = .9492 \text{ month}^{-1}$

(b) $\approx$ twice monthly

Chapter 9

9.1. 216.0 −4.32

135.1 7.04

9.2. $([\mathbf{X}]^T\mathbf{y})^T = [627, 11475]$, $[\mathbf{X}^T\mathbf{X}]^{-1}$

$$= \begin{bmatrix} 0.06263 & -0.002336 \\ -0.002336 & 0.0001797 \end{bmatrix} \mathbf{b}^T = [12.46, 0.60]$$

9.3. 90°

9.4. (a) $\lambda_1 = 118.8$, $\lambda_2 = 2.60$

(b) $\theta = 45.6°$, $2.81\sqrt{\lambda_k}$, $5.97\sqrt{\lambda_k}$

9.5. (a) 2.23, 4.23

(b) 118.8

(c) 0.979

9.6. (a) Correlation matrix: $\Sigma\, \lambda_k = 3$

(b) 1, 3

(c) 95th percentile of $\lambda_1 = 1.66$, $\lambda_2 = 1.12$, $\lambda_3 = 0.86$; retain first principal component

9.7. January 6: $v_1 = 32.7$, $w_1 = 41.1$; January 7: $v_1 = 41.3$, $w_1 = 50.2$

9.8. (a) $\mathbf{d}_1^T = [0.83, -0.56]$

(b) 1953

References

Abramowitz, M., and I. A. Stegun, eds., 1984: *Pocketbook of Mathematical Functions.* Verlag Harri Deutsch, Frankfurt, 468 pp.

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control,* **AC-19**, 716–723.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. N. Rogers, and J. W. Tukey, 1972: *Robust Estimates of Location—Survey and Advances.* Princeton University Press, Princeton, NJ., 373 pp.

Arkin, P. A., 1989: The global climate for December 1988–February 1989: Cold episode in the tropical pacific continues. *J. Clim.*, **2**, 737–757.

Atkinson, K. F., 1978: *An Introduction to Numerical Analysis.* Wiley, New York, 587 pp.

Azcarraga, R., and A. J. Ballester G., 1991: Statistical system for forecasting in Spain. In *Programme on Short- and Medium-Range Weather Prediction Research,* H. R. Glahn, A. H. Murphy, L. J. Wilson, and J. S. Jensenius, Jr., eds., WM/TD No. 421. World Meteorological Organization, Geneva, XX23-25.

Barnett, T. P., and R. W. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Weather Rev.*, **115**, 1825–1850.

Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Clim.*, **5**, 1316–1345.

Baughman, R. G., D. M. Fuquay, and P. W. Mielke, Jr., 1976: Statistical analysis of a randomized lightning modification experiment. *J. Appl. Meteorol.*, **15**, 790–794.

Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *Mon. Weather Rev.*, **97**, 163–172.

Blackmon, M. L., 1976: A climatological spectral study of the 500 mb geopotential height of the northern hemisphere. *J. Atmos. Sci.*, **33**, 1607–1623.

Boden, T. A., P. Kanciruk, and M. P. Farrell, 1990: *Trends '90, A Compendium of Data on Global Change.* Carbon Dioxide Information Analysis Center, Oak Ridge, TN, 257 pp.

Box, G. E. P, and D. R. Cox, 1964: An analysis of transformations. *J. R. Stat. Soc.*, **B26**, 211–243.

Box, G. E. P., and G. M. Jenkins, 1976: *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, 575 pp.

Bras, R. L., and I. Rodríguez-Iturbe, 1985: *Random Functions and Hydrology.* Addison-Wesley, Reading, MA, 559 pp.

Bratley, P., B. G. Fox, and L. E. Schrage, 1983: *A Guide to Simulation.* Springer-Verlag, Berlin, 383 pp.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Clim.*, **5**, 541–560.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Weather Rev.*, **78**, 1–3.

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. In *Compendium of Meteorology*, T. F. Malone, ed., American Meteorological Society, pp. 841–848.

Brooks, C. E. P, and N. Carruthers, 1953: *Handbook of Statistical Methods in Meteorology*. Her Majesty's Stationery Office, London, 412 pp.

Brown, B. G., and R. W. Katz, 1991: Use of statistical methods in the search for teleconnections: Past, present, and future. In *Teleconnections Linking Worldwide Climate Anomalies*, M. Glantz, R. W. Katz, and N. Nicholls, eds., Cambridge University Press, Cambridge, UK, 371–400.

Brunet, N., R. Verret, and N. Yacowar, 1988: An objective comparison of model output statistics and "perfect prog" systems in producing numerical weather element forecasts. *Weather Forecasting*, **3**, 273–283.

Buishand, T. A., 1989: Statistics of extremes in climatology. *Stat. Neerl.*, **43**, 1–30.

Carter, G. M., and P. D. Polger, 1986: *A 20-Year Summary of National Weather Service Verification Results for Temperature and Precipitation*. NOAA Tech. Memo. NWS FCST 31. National Weather Service, Camp Springs, MD, 50 pp.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Weather and Forecasting*, **4**, 401–412.

Cayan, D. R., C. F. Ropelewski, and T. R. Karl, 1986: *An Atlas of United States Monthly and Seasonal Temperature Anomalies, December 1930–November 1984*. Climate Analysis Center, Washington, DC, 244 pp.

Chen, W. Y., 1982a: Assessment of southern oscillation sea-level pressure indices. *Mon. Weather Rev.*, **110**, 800–807.

Chen, W. Y., 1982b: Fluctuations in northern hemisphere 700 mb height field associated with the southern oscillation. *Mon. Weather Rev.*, **110**, 808–823.

Chu, P.-S., and R. W. Katz, 1989: Spectral estimation from time series models with relevance to the southern oscillation. *J. Clim.*, **2**, 86–90.

Conradsen, K., L. B. Nielsen, and L. P. Prahm, 1984: Review of Weibull statistics for estimation of wind speed distributions. *J. Clim. Appl. Meteorol.*, **23**, 1173–1183.

Conte, M., C. DeSimone, and C. Finizio, 1980: Post-processing of numerical models: Forecasting the maximum temperature at Milano Linate. *Rev. Meteorol. Aeronaut.*, **40**, 247–265.

Crutcher, H. L., 1975: A note on the possible misuse of the Kolmogorov-Smirnov test. *J. Appl. Meteorol.*, **14**, 1600–1603.

Daan, H., 1985: Sensitivity of verification scores to the classification of the predictand. *Mon. Weather Rev.*, **113**, 1384–1392.

Dallavalle, J. P., J. S. Jensenius, Jr., and S. A. Gilbert, 1992: *NGM-based MOS Guidance—the FOUS14/FWC Message*, Tech. Proced. Bull. 408. NOAA/National Weather Service, Washington, DC, 9 pp.

Daniel, W. W., 1990: *Applied Nonparametric Statistics*. PWS—Kent, Boston, 635 pp.

Davis, R. E., 1976: Predictability of sea level pressure anomalies over the north Pacific Ocean. *J. Phys. Oceanogr.*, **6**, 249–266.

DeGaetano, A. T., and M. D. Shulman, 1990: A climatic classification of plant hardiness in the United States and Canada. *Agric. For. Meteorol.*, **51**, 333–351.

Diaconis, P., and B. Efron, 1983: Computer-intensive methods in statistics. *Sci. Am.*, **248**, 116–130.

Dixon, W. J., and F. J. Massey, Jr., 1983: *Introduction to Statistical Analysis*. McGraw-Hill, New York, 678 pp.

Downton, M. W., and R. W. Katz, 1993: A test for inhomogeneous variance in time-averaged temperature data. *J. Clim.*, **6**, 2448–2464.

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis*. Wiley, New York, 709 pp.

Efron, B., 1982: *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Providence, RI, 92 pp.

Efron, B., 1987: Better bootstrap confidence intervals. *J. Am. Stat. Assoc.*, **82**, 171–185.

Efron, B., and G. Gong, 1983: A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.*, **37**, 36–48.

Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Weather Rev.*, **122**, 703–713.

Ehrendorfer, M., and A. H. Murphy, 1988: Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy. *Mon. Weather Rev.*, **116**, 1757–1770.

Emerson, J. D., and D. C. Hoaglin, 1983: Stem-and-leaf displays. In *Understanding Robust and Exploratory Data Analysis*. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., Wiley, New York, pp. 7–32.

Emerson, J. D., and M. A. Stoto, 1983: Transforming data. In *Understanding Robust and Exploratory Data Analysis*. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., Wiley, New York, pp. 97–128.

Epstein, E. S., 1969a: The role of initial uncertainties in prediction. *J. Appl. Meteorol.*, **8**, 190–198.

Epstein, E. S., 1969b: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.*, **8**, 985–987.

Epstein, E. S., 1969c: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Epstein, E. S., 1985: Statistical inference and prediction in climatology: A Bayesian approach. *Meteorol. Monogr.*, **20**(42), 199 pp.

Epstein, E. S., 1988: Long-range prediction: Limits of predictability and beyond. *Weather and Forecasting*, **3**, 69–75.

Epstein, E. S., 1991: On obtaining daily climatological values from monthly means. *J. Clim.*, **4**, 365–368.

Epstein, E. S., and A. G. Barnston, 1988: *A Precipitation Climatology of Five-Day Periods*, NOAA Tech. Rep. NWS 41. Climate Analysis Center, National Weather Service, Camp Springs, MD, 162 pp.

Epstein, E. S., and R. J. Fleming, 1971: Depicting stochastic dynamic forecasts. *J. Atmos. Sci.*, **28**, 500–511.

Erickson, M. C., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, E. Jacks, J. S. Jensenius Jr., and J. C. Su, 1991: Evaluating the impact of RAFS changes on the NGM-based MOS guidance. *Weather and Forecasting*, **6**, 142–147.

Feller, W., 1970: *An Introduction to Probability Theory and its Applications*. Wiley, New York, 509 pp.

Finley, J. P., 1884: Tornado prediction. *Am. Meteorol. J.*, **1**, 85–88.

Flueck, J. A., 1987: A study of some measures of forecast verification. *Preprints, Tenth Conference on Probability and Statistics in Atmospheric Sciences*. American Meteorological Society, pp. 69–73.

Fovell, R. G., and M.-Y. Fovell, 1993: Climate zones of the conterminous United States defined using cluster analysis. *J. Clim.*, **6**, 2103–2135.

Francis, P. E., A. P. Day and G. P. Davis, 1982: Automated temperature forecasting, an application of Model Output Statistics to the Meteorological Office numerical weather prediction model. *Meteorol. Mag.*, **111**, 73–87.

Fuller, W. A., 1976: *Introduction to Statistical Time Series*. Wiley, New York, 470 pp.

Galliani, G., and F. Filippini, 1985: Climatic clusters in a small area. *J. Climatol.*, **5**, 487–501.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Weather Rev.*, **120**, 361–370.

Gerrity, J. P., Jr., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Weather Rev.*, **120**, 2709–2712.

Glahn, H. R., 1985: Statistical weather forecasting. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, eds., Westview, Boulder, CO, pp. 289–335.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.*, **11**, 1203–1211.

Gleeson, T. A., 1967: Probability predictions of geostrophic winds. *J. Appl. Meteorol.*, **6**, 355–359.

Goldsmith, B. S., 1990: *NWS Verification of Precipitation Type and Snow Amount Forecasts During the AFOS Era*, NOAA Tech. Mem. NWS FCST 33. National Weather Service, Camp Springs, MD, 28 pp.

Goodall, C., 1983: M-Estimators of location: An outline of the theory. In *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., Wiley, New York, pp. 339–403.

Graedel, T. E., and B. Kleiner, 1985: Exploratory analysis of atmospheric data. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, eds., Westview, Boulder, CO, pp. 1–43.

Gray, W. M., 1990: Strong association between West African rainfall and U.S. landfall of intense hurricanes. *Science*, **249**, 1251–1256.

Greenwood, J. A., and D. Durand, 1960: Aids for fitting the gamma distribution by maximum likelihood. *Technometrics*, **2**, 55–65.

Gringorten, I. I., 1967: Verification to determine and measure forecasting skill. *J. Appl. Meteorol.*, **6**, 742–747.

Gumbel, E. J., 1958: *Statistics of Extremes*. Columbia University Press, New York, 375 pp.

Guttman, N. B., 1993: The use of *L*-moments in the determination of regional precipitation climates. *J. Clim.*, **6**, 2309–2325.

Gutzler, D. S., and J. Shukla, 1984: Analogs in the wintertime 500 mb height field. *J. Atmos. Sci.*, **41**, 177–189.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.*, **81**, 2–15.

Harnack, R., M. Cammarata, K. Dixon, J. Lanzante, and J. Harnack, 1985: Summary of U.S. seasonal temperature forecast experiments. *Preprints, Ninth Conference on Probability and Statistics in the Atmospheric Sciences*. American Meteorological Society, pp. 175–178.

Harter, H. L., 1984: Another look at plotting positions. *Communi. Stat. Theory Methods*, **13**, 1613–1633.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geogr. Ann.*, **8**, 310–349.

Hinkley, D., 1977: On quick choice of power transformation. *Appl. Stat.*, **26**, 67–69.

Hoaglin, D. C., 1983: Letter values: A set of selected order statistics. In *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., Wiley, New York, pp. 33–57.

Hoffman, M. S., ed., 1988: *The World Almanac Book of Facts*. Pharos Books, New York, 928 pp.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35A**, 100–118.

Hoke, J. E., N. A. Phillips, G. J. DiMego, J. J. Tuccillo, and J. G. Sela, 1989: The regional analysis and forecast system of the National Meteorological Center. *Weather and Forecasting*, **4**, 323–334.

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and H. Savijärvi, 1980: The performance of a medium range forecast model in winter—impact of physical parameterizations. *Mon. Weather Rev.*, **108**, 1736–1773.

Horel, J. D., 1981: A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field. *Mon. Weather Rev.*, **109**, 2080–2902.

Hosking, J. R. M., 1990: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc.*, **B52**, 105–124.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293.

Iglewicz, B., 1983: Robust scale estimators and confidence intervals for location. In *Understanding*

Robust and Exploratory Data Analysis. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., Wiley, New York, pp. 404–431.

Ivarsson, K.-I., R. Joelsson, E. Liljas, and A. H. Murphy, 1986: Probability forecasting in Sweden: some results of experimental and operational programs at the Swedish Meteorological and Hydrological Institute. *Weather and Forecasting*, **1**, 136–154.

Jenkins, G. M., and D. G. Watts, 1968: *Spectral Analysis and its Applications*. Holden-Day, San Francisco, 523 pp.

Johnson, M. E., 1987: *Multivariate Statistical Simulation*. Wiley, New York, 230 pp.

Johnson, N. L,. and S. Kotz, 1969: *Distributions in Statistics. 1. Discrete Distributions*. Wiley, New York, 328 pp.

Johnson, N. L,. and S. Kotz, 1970a: *Distributions in Statistics. 2. Continuous Univariate Distributions I*. Wiley, New York, 300 pp.

Johnson, N. L,. and S. Kotz, 1970b: *Distributions in Statistics. 3. Continuous Univariate Distributions II*. Wiley, New York, 306 pp.

Johnson, N. L,. and S. Kotz, 1972: *Distributions in Statistics. 4. Continuous Multivariate Distributions*. Wiley, New York, 333 pp.

Johnson, R. A., and D. A. Wichern, 1982: *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 594 pp.

Jolliffe, I. T., 1972: Discarding variables in a principal component analysis. I. Artificial data. *Appl. Stat.*, **21**, 160–173.

Jolliffe, I. T., 1986: *Principal Component Analysis*. Springer, New York, 271 pp.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteorol.*, **14**, 159–163.

Kalkstein, L. S., G. Tan, and J. A. Skindlov, 1987: An evaluation of three clustering procedures for use in synoptic climatological classification. *J. Clim. Appl. Meteorol.*, **26**, 717–730.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Am. Meteorol. Soc.*, **71**, 1410–1428.

Karl, T. R., M. E. Schlesinger, and W. C. Wang, 1989: A method of relating general circulation model simulated climate to the observed local climate. Part I: Central tendencies and dispersion. *Preprints, Sixth Conference on Applied Climatology*. American Meteorological Society, pp. 188–196.

Karlin, S., and H. M Taylor, 1975: *A First Course in Stochastic Processes*. Academic Press, New York, 557 pp.

Katz, R. W., 1977: Precipitation as a chain-dependent process. *J. Appl. Meteorol.*, **16**, 671–676.

Katz, R. W., 1981: On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 243–249.

Katz, R. W., 1982: Statistical evaluation of climate experiments with general circulation models: A parametric time series modeling approach. *J. Atmos. Sci.*, **39**, 1446–1455.

Katz, R. W., 1985: Probabilistic models. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, eds., Westview, Boulder, CO, pp. 261–288.

Katz, R. W., and M. B. Parlange, 1993: Effects of an index of atmospheric circulation on stochastic properties of precipitation. *Water Resour. Res.*, **29**, 2335–2344.

Katz, R. W., A. H. Murphy, and R. L. Winkler, 1982: Assessing the value of frost forecasts to orchardists: A dynamic decision-making approach. *J. Appl. Meteorol.*, **21**, 518–531.

Kendall, M., and J. K. Ord, 1990: *Time Series*. Edward Arnold, London, p. 296.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperature during winter. *J. Meteorol.*, **16**, 672–682.

Krzysztofowicz, R., 1983: Why should a forecaster and a decision maker use Bayes' theorem? *Water Resour. Res.*, **19**, 327–336.

Krzysztofowicz, R., and D. Long, 1990: Fusion of detection probabilities and comparison of multisensor systems. *IEEE Trans. Syst., Man, Cyberne.*, **SMC-20**, 665–677.

Krzysztofowicz, R., and D. Long, 1991: Beta probability models of probabilistic forecasts. *Int. J. Forecasting*, **7**, 47–55.

Krzysztofowicz, R., W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Weather and Forecasting*, **8**, 424–439.

Kutzbach, J. E., 1967: Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. *J. Appl. Meteorol.*, **6**, 791–802.

Lawson, M. P., and R. S. Cerveny, 1985: Seasonal temperature forecasts as products of antecedent linear and spatial temperature arrays. *J. Clim. Appl. Meteorol.*, **24**, 848–859.

Leith, C. E., 1973: The standard error of time-average estimates of climatic means. *J. Appl. Meteorol.*, **12**, 1066–1069.

Leith, C. E., 1974: Theoretical skill of Monte-Carlo forecasts. *Mon. Weather Rev.*, **102**, 409–418.

Lemcke, C., and S. Kruizinga, 1988: Model output statistics forecasts: Three years of operational experience in the Netherlands. *Mon. Weather Rev.*, **116**, 1077–1090.

Lindgren, B. W., 1976: *Statistical Theory*. Macmillan, New York, 614 pp.

Lipschutz, S., 1968: *Schaum's Outline of Theory and Problems of Linear Algebra*. McGraw-Hill, New York, 334 pp.

Livezey, R. E., 1985: Statistical analysis of general circulation model climate simulation: Sensitivity and prediction experiments. *J. Atmos. Sci.*, **42**, 1139–1149.

Livezey, R. E., and A. G. Barnston, 1988: An operational multifield analog prediction system for the United States seasonal temperatures. Part I: System design and winter experiments. *J. Geophys. Res.*, **93**, 10953–10974.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Weather Rev.*, **111**, 46–59.

Lorenz, E. N., 1956: *Empirical Orthogonal Functions and Statistical Weather Prediction*, Sci. Rep. 1. Statistical Forecasting Project, Department of Meteorology, MIT (NTIS AD 110268), 49 pp.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646.

Lorenz, E. N., 1993: *The Essence of Chaos*. University of Washington Press, Seattle, 227 pp.

Lu, R., 1991: The application of NWP products and progress of interpretation techniques in China. In *Programme on Short- and Medium-Range Weather Prediction Research*, H. R. Glahn, A. H. Murphy, L. J. Wilson, and J. S. Jensenius, Jr., eds., WM/TD No. 421. World Meteorological Organization, Geneva, XX19-22.

Madden, R. A., 1979: A simple approximation for the variance of meteorological time averages. *J. Appl. Meteorol.*, **18**, 703–706.

Madden, R. A., and D. J. Shea, 1978: Estimates of the natural variability of time-averaged temperatures over the United States. *Mon. Weather Rev.*, **106**, 1695–1703.

Mason, S.J., and G. M. Mimmack, 1992: The use of bootstrap confidence intervals for the correlation coefficient in climatology. *Theor. Appl. Climatol.*, **45**, 229–233.

McGill, R., J. W. Tukey, and W. A. Larsen, 1978: Variations of boxplots. *Am. Stat.*, **32**, 12–16.

Michaelson, J., 1987: Cross-validation in statistical climate forecast models. *J. Clim. Appl. Meteorol.*, **26**, 1589–1600.

Mielke, P. W., Jr., K. J. Berry, and G. W. Brier, 1981: Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. *Mon. Weather Rev.*, **109**, 120–126.

Miller, B. I., E. C. Hill, and P. P. Chase, 1968: A revised technique for forecasting hurricane movement by statistical methods. *Mon. Weather Rev.*, **96**, 540–548.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. *Meteorol. Monog.*, **4**, No. 25, 53 pp.

Mo, K. C., and M. Ghil, 1987: Statistics and dynamics of persistent anomalies. *J. Atmos. Sci.*, **44**, 877–901.

Mo, K. C., and M. Ghil, 1988: Cluster analysis of multiple planetary flow regimes. *J. Geophys. Res.*, **D93**, 10927–10952.

Molteni, F., S. Tibaldi, and T. N. Palmer, 1990: Regimes in wintertime circulation over northern extratropics. I: Observational evidence. *Q. J. R. Meteorol. Soc.*, **116**, 31–67.

Morrison, D. F., 1976: *Multivariate Statistical Methods*. McGraw-Hill, New York, 415 pp.

Muller, R. H., 1944: Verification of short-range weather forecasts (a survey of the literature). *Bull. Am. Meteorol. Soc.*, **25**, 18–27, 47–53, 88–95.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Q. J. R. Meteorol. Soc.*, **119**, 299–323.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteorol.*, **10**, 155–156.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteorol.*, **12**, 595–600.

Murphy, A. H., 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Weather Rev.*, **105**, 803–816.

Murphy, A. H., 1985: Probabilistic weather forecasting. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, eds., Westview, Boulder, CO, pp. 337–377.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Weather Rev.*, **116**, 2417–2424.

Murphy, A. H., 1991: Forecast verification: its complexity and dimensionality. *Mon. Weather Rev.*, **119**, 1590–1601.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–293.

Murphy, A. H., and B. G. Brown, 1983: Forecast terminology: Composition and interpretation of public weather forecasts. *Bull. Am. Meteorol. Soc.*, **64**, 13–22.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, eds., Westview, Boulder, CO, pp. 379–437.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Weather Rev.*, **117**, 572–581.

Murphy, A. H., and R. L. Winkler, 1974: Credible interval temperature forecasting: Some experimental results. *Mon. Weather Rev.*, **102**, 784–794.

Murphy, A. H., and R. L. Winkler, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Am. Meteorol. Soc.*, **60**, 12–19.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Weather Rev.*, **115**, 1330–1338.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435–455.

Murphy, A. H., B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485–501.

Murphy, J. M., 1990: Assessment of the practical utility of extended-range ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **116**, 89–125.

National Bureau of Standards, 1959: *Tables of the Bivariate Normal Distribution Function and Related Functions*, Appl. Math. Ser. 50. U.S. Government Printing Office, Washington, DC, 258 pp.

Neter, J., W. Wasserman, and M. H. Kutner, 1985: *Applied Linear Statistical Models*. Irwin, Homewood, IL, 1127 pp.

Neumann, C. J., M. B. Lawrence, and E. L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. *J. Appl. Meteorol.*, **16**, 1165–1174.

Nicholls, N., 1987: The use of canonical correlation to study teleconnections. *Mon. Weather Rev.*, **115**, 393–399.

Overland, J. E., and R. W. Preisendorfer, 1982: A significance test for principal components applied to a cyclone climatology. *Mon. Weather Rev.*, **110**, 1–4.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Am. Meteorol. Soc.*, **74**, 49–65.

Panofsky, H. A., and G. W. Brier, 1958: *Some Applications of Statistics to Meteorology*. Pennsylvania State University, University Park, 224 pp.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Peixoto, J. P., and A. H. Oort, 1992: *Physics of Climate*. American Institute of Physics, New York, 520 pp.

Peterson, C. R., K. J. Snapper, and A. H. Murphy 1972: Credible interval temperature forecasts. *Bull. Am. Meteorol. Soc.*, **53**, 966–970.

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography*. C. D. Mobley, ed., Elsevier, Amsterdam, 425 pp.

Preisendorfer, R. W., and T. P. Barnett, 1983: Numerical-reality intercomparison tests using small-sample statistics. *J. Atmos. Sci.*, **40**, 1884–1896.

Preisendorfer, R. W., and C. D. Mobley, 1984: Climate forecast verifications, United States Mainland, 1974–83. *Mon. Weather Rev.*, **112**, 809–825.

Preisendorfer, R. W., F. W. Zwiers, and T. P. Barnett, 1981: *Foundations of Principal Component Selection Rules*, SIO Ref. Ser. 81-4. Scripps Institution of Oceanography, La Jolla, CA, 192 pp.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1986: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 818 pp.

Quayle, R. and W. Presnell, 1991: *Climatic Averages and Extremes for U.S. Cities*, Hist. Climatol. Ser. 6-3. National Climatic Data Center, Asheville, NC, 270 pp.

Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour. Res.*, **17**, 182–190.

Richman, M. B., 1986: Rotation of principal components. *J. Climatol.*, **6**, 293–335.

Romesburg, H. C., 1984: *Cluster Analysis for Researchers*. Wadsworth/Lifetime Learning Publications, Belmont, CA, 334 pp.

Ropelewski, C. F., and P. D. Jones, 1987: An extension of the Tahiti-Darwin Southern Oscillation index. *Mon. Weather Rev.*, **115**, 2161–2165.

Ropelewski, C. F., J. E. Janowiak, and M. S. Halpert, 1985: The analysis and display of real time surface climate data. *Mon. Weather Rev.*, **113**, 1101–1106.

Rosenberger, J. L., and M. Gasko, 1983: Comparing location estimators: Trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., Wiley, New York, pp. 297–338.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Sheets, R. C., 1990: The National Hurricane Center—past, present and future. *Weather and Forecasting*, **5**, 185–232.

Solow, A. R., 1985: Bootstrapping correlated data. *Math. Geol.*, **17**, 769–775.

Spetzler, C. S., and C.-A. S. Staël von Holstein, 1975: Probability encoding in decision analysis. *Manage. Sci.*, **22**, 340–358.

Staël von Holstein, C-.A. S., and A. H. Murphy, 1978: The family of quadratic scoring rules. *Mon. Weather Rev.*, **106**, 917–924.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: *Survey of Common Verification Methods in Meteorology*, World Weather Watch Tech. Rep. No. 8, TD No. 358. World Meteorological Organization, Geneva, 114 pp.

Stern, R. D., and R. Coe, 1984: A model fitting analysis of daily rainfall data. *J. R. Stat. Soc.*, **A147**, 1–34.

Stull, R. B., 1988: *An Introduction to Boundary Layer Meteorology*. Kluwer, New York, 666 pp.

Thiébaux, H. J., and M. A. Pedder, 1987: *Spatial Objective Analysis : With Applications in Atmospheric Science*. Academic Press, London, 299 pp.

Thom, H. C. S., 1958: A note on the gamma distribution. *Mon. Weather Rev.*, **86**, 117–122.

Thompson, J. C., 1962: Economic gains from scientific advances and operational improvements in meteorological prediction. *J. Appl. Meteorol.*, **1**, 13–17.

Tong, H., 1975: Determination of the order of a Markov chain by Akaike's Information Criterion. *J. Appl. Probability*, **12**, 488–497.

Toth, Z., 1991: Intercomparison of circulation similarity measures. *Mon. Weather Rev.*, **119**, 55–64.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.*, **74**, 2317–2330.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather and Forecasting*, **8**, 379–398.

Tracton, M. S., K. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Weather Rev.*, **117**, 1604–1635.

Tukey, J. W., 1977: *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 688 pp.

Velleman, P. F., 1988: *Data Desk*. Data Description, Inc., Ithaca, NY.

Velleman, P. F., and D. C. Hoaglin, 1981: *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press, Boston, 354 pp.

Wadsworth, G. P., 1951: Application of statistical methods to weather forecasting. In *Compendium of Meteorology*, T. F. Malone, ed., American Meteorological Society, pp. 849–855.

Wagner, A. J., 1989: Medium- and long-range forecasting. *Weather and Forecasting*, **4**, 413–426.

Wallace, J. M., and M. L. Blackmon, 1983: Observations of low-frequency atmospheric variability. In *Large-Scale Dynamical Processes in the Atmosphere*, B. J. Hoskins and R. P. Pearce, eds., Academic Press, New York, pp. 55–94.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather Rev.*, **109**, 784–812.

Wallace, J. M., C. Smith, and C. S. Bretherton, 1992: Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *J. Clim.*, **5**, 561–576.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.*, **11**, 711–743.

Waymire, E., and V. K. Gupta, 1981: The mathematical structure of rainfall representations. 1. A review of stochastic rainfall models. *Water Resour. Res.*, **17**, 1261–1272.

Wigley, T. M. L, and B. D. Santer, 1990: Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.*, **D95**, 851–865.

Wilks, D. S., 1989: Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resour. Res.*, **25**, 1429–1439.

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Clim.*, **3**, 1495–1501.

Wilks, D. S., 1992: Adapting stochastic weather generation algorithms for climate change studies. *Clim. Change*, **22**, 67–84.

Wilks, D. S., 1993: Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series. *Water Resour. Res.*, **29**, 3543–3549.

Wilks, D. S., 1995: Forecast value: prescriptive decision studies. In *Economic Value of Weather and Climate Forecasts*, R. W. Katz and A. H. Murphy, eds., Cambridge University Press, Cambridge, UK (in press).

Wilks, D. S., and K. L. Eggleson, 1992: Estimating monthly and seasonal precipitation distributions using the 30- and 90-day outlooks. *J. Clim.*, **5**, 252–259.

Winkler, R. L., 1972: *Introduction to Bayesian Inference and Decision*. Holt, Rinehart & Winston, New York, 563 pp.

Winkler, R. L., and A. H. Murphy, 1979: The use of probabilities in forecasts of maximum and minimum temperatures. *Meteorol. Mag.*, **108**, 317–329.

Wolter, K., 1987: The southern oscillation in surface circulation and climate over the tropical Atlantic, eastern Pacific, and Indian Oceans as captured by cluster analysis. *J. Clim. Appl. Meteorol.*, **26**, 540–558.

Wong, R. K. W., 1977: Weibull distribution, iterative likelihood techniques and hydrometeorological data. *J. Appl. Meteorol.*, **16**, 1360–1364.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Weather Rev.*, **104**, 1209–1214.

Woolhiser, D. A., and J. Roldan, 1982: Stochastic daily precipitation models. 2. A comparison of distributions of amounts. *Water Resour. Res.*, **18**, 1461–1468.

Zeng, X., R. A. Pielke and R. Eykhold, 1993: Chaos theory and its application to the atmosphere. *Bull. Am. Meteorol. Soc.*, **74**, 631–644.

Zwiers, F. W., 1987: Statistical considerations for climate experiments. Part II: Multivariate tests. *J. Clim. Appl. Meteorol.*, **26**, 477–487.

Zwiers, F. W., 1990: The effect of serial correlation on statistical inferences made with resampling procedures. *J. Clim.*, **3**, 1452–1461.

Zwiers, F. W., and H. J. Thiébaux, 1987: Statistical considerations for climate experiments. Part I: Scalar tests. *J. Clim. Appl. Meteorol.*, **26**, 465–476.

This Page Intentionally Left Blank

Index

AC, *see* Anomoly correlation
Accuracy, 236
Additive law of probability, 11–12
AIC, *see* Akaike information criterion
Akaike information criterion, 300–301, 313
Aliasing, 347–350
Alternative hypothesis, 115
Amplitude, 327, 396
Analog forecasting, 198–199
Analysis formula, principal component analysis, 376, 379, 384, 397, 398
Analysis of variance, 165–166
Anomaly, 42, 86, 285, 304, 308, 374, 407
Anomaly correlation, 277–281
ANOVA, *see* Analysis of variance
ARMA, *see* Autoregressive moving-average model
AR(1) model, 302–308, 314, 316, 318, 320–321, 324, 352–354
AR(2) model, 309–313, 320–321, 322, 354
Artificial Skill, 191
Atmospheric field, 51, 61, 198, 210, 272–281, 359, 374, 386–393, 396, 399, 403–408
Attributes diagram, 263–265
Autocorrelation, 51–53, 126–127, 174, 291–292, 305, 324, 383
Autocorrelation function, 53–54, 59, 284, 286, 292, 305, 308, 310–312, 315, 318, 354
Autoregressive model, 302–325
  first-order autoregression, 302–308
  forecasting with, 319–323
  higher-order autoregression, 308–314
  multivariate autoregression, 323–325
  simulation with, 319–323
  theoretical spectra, 352–357
  variance of time average, 315–318
Autoregressive moving-average model, 318–319
Average-linkage clustering, 421
Axioms of probability, 8

Backward elimination, 189
Bayesian information criterion, 300–302, 313–314
Bayesian interpretation of probability, 9–10
Bayes' theorem, 17–19
Bernoulli distribution, 69, 291
Beta distribution, 95–97
Between-groups variance matrix, 416
BC, *see* Bias-corrected confidence interval
Bias, 110, 200, 205, 228, 236, 241–243, 247, 254, 256, 266, 275, 278
Bias-corrected confidence interval, 148
Bias ratio, 241
BIC, *see* Bayesian information criterion
Binary forecasts, accuracy measures, 239–241
Binary variable, 178
Binomial distribution, 66–70, 74–76, 119–121, 152–153, 288, 291–292, 300–301
Bivariate normal distribution, 99–102
  confidence ellipse, 372–373
Blue noise, 354
Bootstrap, 145–148, 150, 383
Bootstrap confidence interval, 146–148
Box-and-whisker plot, 29–30
Box–Cox transformation, 37
Box–Jenkins models, 302
Boxplot, 29–30
  notched, 33
  variant, 33
Brier score, 259–260
  algebraic decomposition of, 260–263
Broken stick model, 382
Brushing, 58–59
BS, *see Brier score*

Calibration-refinement factorization, 235
Canonical correlation analysis, 398–399
  canonical variates, 399–400
  computation, 400–403
  fields, 403–407
  principal-component analysis, 408
Canonical pattern, 403–405
Canonical variate, 399–400, 417
Canonical vectors, 400–403
Categorical forecast
  accuracy measures, 239–244

contingency table, 238–239
conversion from probabilistic, 244–248
discrete predictands, 238–250
fields, 273–281
mean-squared error, 275–277
S1 score, 274–275
skill score, 248–250
CCA, *see* Canonical correlation analysis
CDF, *see* cumulative distribution function
Central credible-interval forecast, 226–227
Central limit theorem, 79, 84, 120–123, 128
Centroid clustering, 421
Chain-dependent process, 296
Chaining, 422
Chaos, dynamical, 3, 210
Characteristic vector, 370
Chi-square distribution, 134, 136–137, 294, 350, 351, 372, 438
Chi-square test, goodness-of-fit, 133–135, 298, 301
Classical nonparametric test, 114, 138–145
Wilcoxon–Mann–Whitney test, 138–142
Wilcoxon signed-rank test, 138, 142–145
Classical statistical forecasting, 159, 181–182
compared to perfect Prog and MOS, 202–209
Climatology
conditional, 14–15
sample, 235
CLT, *see* Central limit theorem
Cluster analysis, 419–420
clustering methods, 420–424
dendogram, 423–424
distance measures, 420–424
stopping level, 424–427
Coefficient of determination, 167
Collectively exhaustive, 8
Complementary event, 11
Complete-linkage clustering, 421
Compositing, 182
Conditional bias, 236, 256, 262
Conditional climatology, 14–15
Conditional probability, 12–18
bivariate normal distribution, 101–102
Conditional quantile plot, 251–252
Conditioning event, 13
Confidence ellipse, 372–373
Confidence interval, 118–121
bias-corrected, 148
bootstrap, 146–148
prediction, 175–176
Contingency table, 238–239
marginal totals, 239
skill measures for, 248–250
Continuous distribution, 66
beta distribution, 95–97
functions and expected value, 76–79
gamma distribution, 86–93, 105, 107, 110–111, 131–132, 134–137, 294, 296, 350, 351, 372, 435
Gaussian distribution, 22, 42, 44, 58, 65, 66, 79–86, 88, 94, 95, 99–103, 105–108, 109–110, 120, 121–123, 128, 130–135, 140, 154, 163, 168, 172, 175, 303, 321
Gumbel distribution, 97–98
lognormal distribution, 95
Weibull distribution, 93–94
Continuous random variable, 78–79
Control forecast, 237
Correlation
auto, 52
lagged, 52
Pearson, 45–50
rank, 50–51
serial, 51–53
Spearman rank, 50–51
Correlation coefficient, 45
Correlation map, 59–62
one point, 59–60
Correlation matrix, 55–57, 366–368
principal-component analysis, 383–386
Cosine function, 325–326
Counting norm statistic, 153
Covariance, 46–47, 103, 124, 367–368, 374
Covariance matrix, principal-component analysis, 383–386
Covariance stationarity, 284
Credible interval, 118
Credible interval forecast, 225–227
Critical region, 116–117, 132
Critical success index, 240
Cross-validation, 194–198
CSI, *see* Critical success index
Cumulative distribution function, 34–36, 78
Cumulative frequency distribution, 34–36

Deciles, 23
Degrees of freedom, 121–122, 235
Dendogram, 423–424
Dependent sample, 186, 194
Dependent variable, 15–16, 160

Derived predictor variables, 177–181
Descriptive statistics, 2
Developmental sample, 186
Diagnostic verification, 234
Diagonal matrix, 367–368
Dichotomous events, 259
  probability forecast, 259
Discrete distribution, 66
  binomial distribution, 66–70, 74–76, 119–121, 152–153, 288, 291–292, 300–301
  geometric distribution, 70–71, 74, 75
  Poisson distribution, 71–73, 74, 75
Discrete Fourier transform, 341–346
Discrete random variable, 286, 287
Discriminant analysis, 408–409
  linear, 409–415
  multiple, 415–419
Discriminant coordinates, 416–417
Discriminant function, 409
Discriminant space, 416
Discrimination, 237
Dispersion, 25
Dispersion matrix, 367
Distribution-free test, 114, 137; *see also* Nonparametric test
Distribution parameters, 65–66
Dominant variance selection rule, 380–382
Dot product, 361
Dummy variable, 183
Durbin–Watson test, 173–174
Dynamical chaos, 3, 310

EDA, *see* Exploratory data analysis
Eddy covariance, 46
Effective sample size, 127, 128, 316, 383
Eigenvalue, 369–373, 376, 380, 396, 401, 416
Eigenvector, 369–373, 375–376, 387, 396, 401, 416
  rotation, 393–394
El Niño, 44, 60, 345, 431
El Niño–Southern Oscillation, 43–44, 60, 345, 431
Empirical cumulative distribution function, 34–36
Empirical orthogonal function analysis, *see* Principal-component analysis
Empirical orthogonal variable, 396
Empirical orthogonal weights, 396
Ensemble average, 216–217
Ensemble forecast, 210–221
  averaging, 216–217
  dispersion, 217–221
  initial conditions, 213–214
  member selection, 215–216
  stochastic dynamic prediction, 211–213
ENSO, *see* El Niño–Southern Oscillation
Equitability, 250
Equivalence of events, 240
Equivalent number of independent samples, 127
Error bar, 118
Error type, 116–117
Euclidean norm, 362
Event, 7
  collectively exhaustive, 8
  complementary, 11
  conditioning, 13
  dependence, 15–16
  independent, 14
  mutually exclusive, 8
  nominal, 268
  null, 7
  ordinal, 268
  subset, 10–11
  union, 11
Event space, 7–8
Expansion coefficient, 396
Expected value, 74–76, 78
Exploratory data analysis, 21
  graphical summary techniques, 27–36
  higher-dimensional data, 54–62
  numerical summary measures, 24–27
  paired data, 44–54
  quantiles, 22–24
  reexpression, 36–44
  resistance, 21–22
  robustness, 21–22
Exponential distribution, 88, 94
Extreme event, 97–98
Extreme-value distribution, 98

Factor analysis, 373
FAR, *see* False-alarm rate
False-alarm rate, 240–241
Fast Fourier transforms, 346–347
Fences, 31
FFTs, *see* Fast Fourier transforms
Field significance, 151–157
First-order autoregression model, 302–308
Fisher's linear discriminant function, 410
Fisher–Tippett type I distribution, 98

Fisher $\mathcal{Z}$ transformation, 153–155
Folding frequency, 348
Forecast quality, 236–237
Forecast skill, 237–239
Forecast verification, 233–234
  bias, 241–242
  categorical forecasts of continuous predictands, 250–258
  categorical forecasts of discrete predictands, 238–250
  categorical forecasts of fields, 273–281
  factorization of the joint distribution of forecasts and observations, 234–235
  forecast skill, 237–238
  scalar attributes, 235–236
Forward selection, 188
Fourier line spectrum, 341–346
Fractiles, 22
F ratio, 168
Frequency-domain analysis
  harmonic analysis, 325–341
  spectral analysis, 341–357
Frequency interpretation of probability, 9
F test, 168
Full rank, 365
Fundamental frequency, 325, 327

Gamma distribution, 86–93, 105, 107, 110–111, 131–132, 134–137, 294, 296, 350, 351, 372, 435
Gamma function, 86–87
Gaussian approximation to the binomial, 120–121
Gaussian distribution, 22, 42, 44, 58, 65, 66, 79–86, 88, 94, 95, 105–108, 120, 121–123, 128, 130–135, 140, 154, 163, 168, 172, 175, 303, 321
  bivariate normal distribution, 99–102
  multivariate normal distribution, 102–103
  parameter fitting, 109–110
Geometric distribution, 70–71, 74, 75
Global significance, 151
Goodness of fit
  measures, 166–168
  probability plot, 106–108
  qualitative assessment of, 104–108
Goodness-of-fit test, 129–135
  chi-square test, 133–134
  Kolmogorov–Smirnov test, 129–133
  Lilliefors test, 129
  Smirnov test, 133
Graphical summary techniques
  boxplot, 29–30
  boxplot variants, 33
  cumulative frequency distribution, 34–36
  harmonic dial, 336–339
  histogram, 33–34
  schematic plot, 30–33
  stem-and-leaf display, 27–29
Gringorten skill score, 249
Gumbel distribution, 97–98

Half-Brier score, 260
Hanssen–Kuipers discriminant, 249–250
Harmonic analysis, 325–341
  cosine function, 325–326
  estimating amplitude and phase, 330–333
  harmonic dials, 336–339
  higher harmonics, 333–336
  simple time series, 326–329
  sine function, 325–326
  uncorrelated regression predictors, 339–341
Harmonic dial, 336–339
Harmonic function, 326–329
  regression predictor, 339–341
Harmonic predictor, 339–341
Hedging, 267–268
Heidke score, 248–250
Heteroscedasticity, 171
Higher harmonics, 333–336
Hinges, 23
Hinkley d_λ statistic, 39–40
Histogram, 33–34
Hit rate, 240
Homoscedasticity, 172
Hypothesis testing
  confidence intervals, 118–119
  elements of, 115–116
  error types, 116–117
  field significance, 151–157
  level of, 116
  multiplicity, 151–157
  nonparametric tests, 137–150
  one-sided, 117–118
  parametric tests, 121–137
  power of, 116–117
  p value, 116
  sampling distribution, 115

two-sided, 117–118
types, 114–115

Identity matrix, 363–364
Incomplete beta function, 97
Incomplete gamma function, 90
Independence, 14, 122–125
Independent variable, 14
Index variable, 4
Inferential statistics, 2
Inner product, 361
Innovation variance, 305–306
Interquartile range, 25–26
Intersection operator, 11
IQR, *see* Interquartile range

Joint probability, 11–12, 13, 14, 16–17, 211, 293

Kaiser's rule, 382
Kolmogorov–Smirnov test, 129–133
Kuipers' performance index, 249–250

Ladder of powers, 37–38
Lagged-average forecasting, 216
Lagged correlation, 52
Latent vector, 370
Law of large numbers, 9
Law of total probability, 16–18
Least error, 160, 163
Least-squares regression
analysis-of-variance table, 165–166
derived predictor variables, 177–181
distribution of residuals, 163–165
examining residuals, 171–175
goodness-of-fit measures, 166–168
multiple linear regression, 176–177
prediction intervals, 175–176
sampling distributions, 168–171
simple linear regression, 160–162
LEPS, *see* Linear error in probability space
LEV, *see* Log-eigenvalue diagram
Level, test, 116
Likelihood, 19, 108
Likelihood-base rate factorization, 235
Likelihood function, 108–110
Likelihood ratio test, 135–137
Markov chain, 301–302
Lilliefors test, 129
Linear algebra, *see* Matrix algebra
Linear discriminant analysis, 409–415
Linear error in probability space score, 257
Linear regression
multiple, 176–181
simple, 160–162, 169–171
Loading, 396
Local test, 151
Location, 24–25
Log-eigenvalue diagram, 380–381
Logistic regression, 183–185
Log-likelihood, 109
Lognormal distribution, 95
L-scale, 149–150

MAD, *see* Median absolute deviation
MAE, *see* Mean absolute error
Marginal distribution, 235
Markov chain, 287–302
higher-order, 298–302
multistate, 297–298
two-state, first order, 288–297
Markovian property, 287
Markov process, 304
Matrix
correlation, 55–57, 366–368
definition, 362–363
dispersion, 367
identity, 363
scatterplot, 57–59
square, 363
symmetrical, 363
variance-covariance, 367, 415–416
Matrix algebra, 360, 362–373
arithmetic operations, 366
eigenvalues and eigenvectors, 369–373
vectors, 360–362
Matrix inverse, 366
Matrix multiplication, 364–366
Maximum-distance clustering, 421
Maximum likelihood, 108–110
Maximum likelihood estimator, 89, 110–111
Mean
tests for differences of, 122–129
trimmed, 25
Mean absolute error, 251–254

Mean error, *see* Bias
Mean-squared error, 166
  continuous categorical forecasts, 253–256
  field forecasts, 275–277
  stopping criterion, 191–192
MECE, *see* Mutually exclusive and collectively exhaustive
Median, 23–24, 29, 31, 226, 229
Median absolute deviation, 26
Method of maximum liklihood, 108–110
Method of moments, 72, 81, 88, 94, 96, 98, 108
Method of successive subdivision, 229
Minimum-distance clustering, 421
MLE, *see* Maximum likelihood estimator
Model output statistics, 201–210
Modes of variation, 396
Monte-Carlo test, 145
  ensemble forecasting, 210–221
MOS, *see* Model output statistic
MSE, *see* Mean-squared error
Multicategory forecast, 242–244
Multicollinearity, 379
Multiple discriminant analysis, 415–419
Multiple linear regression, 176–181
  matrix notation, 368–369
Multiplicative law of probability, 14
Multiplicity, 151–157
Multivariate autoregression, 323–325
Multivariate data analysis
  canonical correlation analysis, 398–408
  cluster analysis, 419–427
  discriminant analysis, 408–419
  matrix algebra notation, 360–373
  principal-component analysis, 373–398
Multivariate data set, 359
Multivariate normal distribution, 102–103
Multivariate probability distribution, 99–103
Mutually exclusive, 8
Mutually exclusive and collectively exhaustive, 8

Nominal event, 268
Nominal predictand, 268
Nonparametric test, 114–115, 137–150
  classical, 114, 138–145
  resampling, 138, 145–150
Normal distribution, 65, 79; *see also* Gaussian distribution
  bivariate, 99–102
  multivariate, 102–103
Normalization transformation, 42
Normal probability plot, 58
Null distribution, 115–116
Null event, 7
Null hypothesis, 115
Numerical weather prediction models, 159, 199–200
  model output statistics, 201–210
  perfect prog, 200–201, 202–209
NWP, *see* Numerical weather prediction models
Nyquist frequency, 342

Objective forecast
  analog forecasting, 198–199
  compositing data, 182
  cross validation, 194–198
  model output statistics, 201–210
  with numerical weather prediction, 199–210
  without numerical weather prediction, 181–199
  overfitting, 185–188
  perfect prog, 200–201, 202–209
  predictand probability, 182–185
  predictor selection, 185–188
  screening regression, 188–189
  stopping rules, 189–194
  stratifying data, 182
Ogive, 34–36
One-sample *t* test, 121–122
One-sided test, 117–118, 121–122
One-tailed test, 117–118
Orbit, 211
Order statistics, 23
Ordinal event, 268
Ordinal predictand, 268
Orthogonality, 362, 370
Outliers, 22, 31
Overfitting, 185–188
  cross-validation and, 195–196
Overforecasting, 241

Pacific–North America pattern, 61, 388, 394, 405
Paired data, 44
  Pearson correlation, 45–50
  scatterplots, 44–45
Paired samples, 124
Parametric distribution, 65
Parametric test, 114–115
  differences of mean, 122–129

goodness-of-fit, 129–135
likelihood ratio, 135–137
one-sample t, 121–122
Pattern correlation, 199, 278
Pattern significance, 151
Pattern vector, 396
PCA, *see* Principal-component analysis
PDFs, *see* Probability density functions
Pearson correlation, 45–50
Percentage of forecasts correct, 240–241
Percentile, 22
Percentile method, 147–148
Perfect Prog, 200–201
comparison with classical, 202–209
Period, 327
Periodogram, 341–346
Permutation test, 148–150
Persistence, 15–16, 51–52, 125, 127, 150, 197, 222, 237, 255, 276, 278, 280–281, 288, 297, 303, 305
Persistence parameter, 291–292
PFC, *see* Percentage of forecasts correct
Phase angle, 328
Phase shift, 328
Phase space, 211
Plotting positions, 35–36
PNA, *see* Pacific-North America pattern
POD, *see* Probability of detection
Poisson distribution, 71–73, 74, 75
Poisson parameter, 71
Poisson process, 71
Pooled estimate of variance, 124
PoP, *see* Probability of precipitation forecast
Positive dependence, 15
Postagreement, 241
Posterior probability, 19
Power function, 117
Power spectrum, 341
Power of a test, 116–117
Power transformation, 37–41
P-P, *see* Probability plot
Predictand, 160
categorical forecast of discrete, 238–250
probability estimates, 182–185
selection, 185–188
Predictive distribution, 235
Predictor variable, 160, 177–181
Prefigurance, 240
Principal component, 374, 396
Principal-component analysis, 373–378
CCA and, 408
correlation versus covariance, 383–386
field analysis, 386–393
rotation of eigenvectors, 393–394
scaling conventions, 397–398
selection rules, 380–382
terminology, 394–397
truncation of components, 378–380
Principal-component selection rules, 380–382
Principal-components regression, 379–380
Prior probability, 18
Probabilistic field forecast, *see* Ensemble forecast
Probability, 6
additive law, 11–12
axioms of, 8
Bayesian interpretation, 9–10
Bayes' Theorem, 17–19
conditional, 12–18
domain, 10–11
event, 7–8
frequency interpretation, 9
law of total probability, 16–18
likelihood, 19
multiplicative law, 14
posterior, 19
prior, 18
properties, 10–14
sample space, 7–8
Probability density functions, 77
Probability of detection, 240
Probability distribution function
binomial distribution, 68
geometric, 70–71
Poisson, 71–73
Probability forecast, verification of, 258–272
Probability-of-precipitation forecast, 223, 246
Probability plot, 106–108
Probability score
ranked, 269–272
strictly proper, 267–268
Product notation, 5
Projection of a vector, 362
Proper vector, 370
Pseudoperiodicity, 311
p value, 116

Q–Q, *see* Quantile–quantile plot
Quantile function, 78
Quantile-quantile plot, 106–107
Quantiles, 22–24
Quartiles, 23–24

Random, 4
Random-number generator, 146, 295
Random variable
  cumulative distribution function for, 78
  expectations for continuous, 78–79
  expected value of, 74–76
  Gaussian, 80
Rank correlation, 50–51
Ranked data, 23, 50–51
Ranked probability score, 269–272
Rank-sum test, 138–145
Rectangular distribution, 146
Red noise, 304–305, 354
REEP, *see* Regression estimation of event probabilities
Reexpression, 36–37
  power transformation, 37–41
  standardized anomoly, 41–44
Reference forecast, 237
Refinement, 237
Regression, *see* Least-squares regression; Linear regression
Regression coefficient, sampling distribution, 168–171
Regression constant, 177
Regression equation, 161
Regression estimation of event probabilities, 183–185
Regression parameters, 177
Regression sum of squares, 165
Rejection level, 116
Rejection region, 116–117
Reliability, 236, 262
  perfect, 263
Reliability diagram, 265–267
Rerandomization test, 145
Resampling test, 145–150
Residual, 160–161
  distribution of, 163–165
  examining, 171–175
Residual plot, 171–175
Residual variance, 163, 305–306
Resistant statistics, 21–22
Resolution, 236, 262–263
RMSE, *see* Root-mean-square error
Robust statistics, 21–22
Root-mean-square error, 276–277
Rotated principal component, 393–394
RPS, *see* Ranked probability score
Rule *N*, 382
S1 score, 274–275
Sample, 127
Sample climatology, 235
Sample space, 7–8
Sampling distribution, 115
Scale, 25
Scale parameter, 86, 88, 89, 90, 94, 98
Scatterplot, 44–45
Scatterplot matrix, 57–59
Schematic plot, 30–33
Scree graph, 380–381
Screening regression, 188–189
Selection rules
  Akaike information criterion, 300–301, 313
  Bayesian information criterion, 300–302, 313–314
  broken-stick model, 382
  Kaiser's rule, 382
  log-eigenvalue diagram, 380–381
  principal components, 380–383
  Rule *N*, 382
  scree graph, 380–381
Serial correlation, 51–53
Serial dependence, 125–129
Shape parameter, 86–90, 94
Sharpness, 237
Significance testing, 114; *see also* Hypothesis testing
Sine function, 325–326
Single-linkage clustering, 421
Skewness, measures of, 26–27
Skill score, 237–238
  contingency table, 248–250
  continuous predictands, 255–257
  probability forecast, 265
Smirnov test, 133
Spatial correlation, 145, 151, 152–157
Spectral analysis, 341–357
  aliasing, 347–350
  computing spectra, 346–347
  Fourier line spectrum, 341–346
  periodogram, 341–346
  sampling properties of spectral estimates, 350–352
  theoretical spectra of autoregressive models, 352–356
Spectrum, 341
Spread, 24, 25–26
SS, *see* Skill score
SSR, *see* Regression sum of squares

SST, *see* Sum of squares, total
Standard deviation, 25, 26, 42, 46, 49, 53, 66, 80, 95, 98, 99, 110, 121, 368
Standard gamma distribution, 90, 435
Standard Gaussian distribution, 81, 433
Standardized anomaly, 41–44, 81, 86, 101
Stationarity, 284–285
Stationary probability, 291
Station model, meteorological, 54
Statistical expectation, 74–76, 78
Statistical forecasting
  classical, 159
  least-squares regression, 160–181
  objective, 159
  subjective, 159
Stem-and-leaf display, 27–29
Stepwise regression, 188
Stochastic dynamic prediction, 211–213
Stopping criterion, 189–194
Stopping rules, 189–194
Stratification, 182
Strictly proper scoring rule, 267–268
Strict stationarity, 284
Student's *t*, 121–122
Subjective distribution, 223–225
Subjective forecast, 221–230
Summation notation, 4
Sum of squares, total, 164
Symmetry, 24, 26–27
  matrix, 363
Synthesis formula, in principal component analysis, 376
Synthetic time series, 295, 304, 311, 319
Synthetic weather generator, 324
Systematic bias, 236

t distribution, 121–122
Teleconnection, 61
Teleconnectivity
  pattern, 388–390
  value, 61
Temporal autocorrelation, 52
Test level, 116
Test statistic, 115
Theoretical distribution
  assessing goodness of fit, 104–108
  background, 64–66
  continuous distributions, 76–98
  discrete distributions, 66–73
  multivariate distributions, 99–103
  parameter fitting, 108–111
  statistical expectation, 74–76
Thom estimator, 89
Threat score, 240
Time between effectively independent samples, 127
Time-domain analysis, 286
  autoregressive models, 302–325
  continuous data, 302–325
  discrete data, 287–302
  Markov chains, 287–302
Time-series analysis, *see also* Harmonic analysis; Spectral analysis; Time-domain analysis
  stationarity, 284–285
  time-series models, 285–286
Training data, 181–182
Training sample, 186
Trajectory, 211
Transition probability, 287
Transpose, vector, 360–361
t-ratio, 169
Tree-diagram, 423–424
Trimean, 24–25
Trimmed mean, 25
Trimmed variance, 26
True skill statistic, 249–250
TS, *see* Threat score
TSS, *see* True skill statistic
t test
  one-sample, 117–118, 121–122
  two-sample, 117–118, 122–125
Two-sided test, 117–118
Two-tailed test, 117–118
Type I error, 116–117
Type II error, 116–117

Uncertainty, 2–3, 263
Unconditional bias, 236
Underforecast, 241
Undersample, 347
Uniform distribution, 146
Uniform-random-number generator, 146
Union, 11
U-statistic, 138–142

Variable
  dependent, 15–16

independent, 14
index, 4
random, 4, 74–79
Variance, 75, 88, 94, 95, 121, 122, 123
sample, 25, 46
trimmed, 26
Variance-covariance matrix, 103, 367–368, 400–401, 415–416
Variance inflation factor, 127
Variance stabilizing transformation, 37
Variance of a time average, 315–318
Vector, 360–362
Vector addition, 361
Vector length, 362–363
Vector multiplication, 361
Venn diagram, 7–8
Verification data, 187

Waiting distribution, 71
Ward's minimum variance clustering, 421
Weak stationarity, 284
Weibull distribution, 93–94
White noise, 353
White-noise forcing, 354
White-noise variance, 305–306, 309, 354
Wilcoxon–Mann–Whitney test, 138–142
Wilcoxon signed-rank test, 138, 142–145
Within-groups variance matrix, 415–416

x–y plot, 44–50

Yule–Kendall index, 27
Yule–Walker equations, 308–309

$\mathfrak{Z}$-transform, for the correlation coefficient, 153–156

International Geophysics Series

EDITED BY

RENATA DMOWSKA
Division of Applied Sciences
Harvard University
Cambridge, Massachusetts

JAMES R. HOLTON
Department of Atmospheric Sciences
University of Washington
Seattle, Washington

Volume 1 BENO GUTENBERG. Physics of the Earth's Interior. 1959*

Volume 2 JOSEPH W. CHAMBERLAIN. Physics of the Aurora and Airglow. 1961*

Volume 3 S. K. RUNCORN (ed.). Continental Drift. 1962*

Volume 4 C. E. JUNGE. Air Chemistry and Radioactivity. 1963*

Volume 5 ROBERT G. FLEAGLE AND JOOST A. BUSINGER. An Introduction to Atmospheric Physics. 1963*

Volume 6 L. DUFOUR AND R. DEFAY. Thermodynamics of Clouds. 1963*

Volume 7 H. U. ROLL. Physics of the Marine Atmosphere. 1965*

Volume 8 RICHARD A. CRAIG. The Upper Atmosphere: Meteorology and Physics. 1965*

Volume 9 WILLIS L. WEBB. Structure of the Stratosphere and Mesosphere. 1966*

Volume 10 MICHELE CAPUTO. The Gravity Field of the Earth from Classical and Modern Methods. 1967*

Volume 11 S. MATSUSHITA AND WALLACE H. CAMPBELL (eds.). Physics of Geomagnetic Phenomena. (In two volumes.) 1967*

Volume 12 K. YA. KONDRATYEV. Radiation in the Atmosphere. 1969*

Volume 13 E. PALMEN AND C. W. NEWTON. Atmospheric Circulation Systems: Their Structure and Physical Interpretation. 1969*

Volume 14 HENRY RISHBETH AND OWEN K. GARRIOTT. Introduction to Ionospheric Physics. 1969*

*Out of print.

Volume 15 C. S. RAMAGE. Monsoon Meteorology. 1971*

Volume 16 JAMES R. HOLTON. An Introduction to Dynamic Meteorology. 1972*

Volume 17 K. C. YEH AND C. H. LIU. Theory of Ionospheric Waves. 1972*

Volume 18 M. I. BUDYKO. Climate and Life. 1974*

Volume 19 MELVIN E. STERN. Ocean Circulation Physics. 1975

Volume 20 J. A. JACOBS. The Earth's Core. 1975*

Volume 21 DAVID H. MILLER. Water at the Surface of the Earth: An Introduction to Ecosystem Hydrodynamics. 1977

Volume 22 JOSEPH W. CHAMBERLAIN. Theory of Planetary Atmospheres: An Introduction to Their Physics and Chemistry. 1978*

Volume 23 JAMES R. HOLTON. An Introduction to Dynamic Meteorology, Second Edition. 1979*

Volume 24 ARNETT S. DENNIS. Weather Modification by Cloud Seeding. 1980

Volume 25 ROBERT G. FLEAGLE AND JOOST A. BUSINGER. An Introduction to Atmospheric Physics, Second Edition. 1980

Volume 26 KUO-NAN LIOU. An Introduction to Atmospheric Radiation. 1980

Volume 27 DAVID H. MILLER. Energy at the Surface of the Earth: An Introduction to the Energetics of Ecosystems. 1981

Volume 28 HELMUT E. LANDSBERG. The Urban Climate. 1981

Volume 29 M. I. BUDYKO. The Earth's Climate: Past and Future. 1982

Volume 30 ADRIAN E. GILL. Atmosphere to Ocean Dynamics. 1982

Volume 31 PAOLO LANZANO. Deformations of an Elastic Earth. 1982*

Volume 32 RONALD T. MERRILL AND MICHAEL W. MCELHINNY. The Earth's Magnetic Field: Its History, Origin, and Planetary Perspective. 1983

Volume 33 JOHN S. LEWIS AND RONALD G. PRINN. Planets and Their Atmospheres: Origin and Evolution. 1983

Volume 34 ROLF MEISSNER. The Continental Crust: A Geophysical Approach. 1986

Volume 35 M. U. SAGITOV, B. BODRI, V. S. NAZARENKO, AND KH. G. TADZHIDINOV. Lunar Gravimetry. 1986

Volume 36 JOSEPH W. CHAMBERLAIN AND DONALD M. HUNTEN. Theory of Planetary Atmospheres: An Introduction to Their Physics and Chemistry, Second Edition. 1987

Volume 37 J. A. JACOBS. The Earth's Core, Second Edition. 1987

Volume 38 J. R. APEL. Principles of Ocean Physics. 1987

Volume 39 MARTIN A. UMAN. The Lightning Discharge. 1987

Volume 40 DAVID G. ANDREWS, JAMES R. HOLTON, AND CONWAY B. LEOVY. Middle Atmosphere Dynamics. 1987

Volume 41 PETER WARNECK. Chemistry of the Natural Atmosphere. 1988

Volume 42 S. PAL ARYA. Introduction to Micrometeorology. 1988

Volume 43 MICHAEL C. KELLEY. The Earth's Ionosphere. 1989

*Out of print.

Volume 44 WILLIAM R. COTTON AND RICHARD A. ANTHES. Storm and Cloud Dynamics. 1989

Volume 45 WILLIAM MENKE. Geophysical Data Analysis: Discrete Inverse Theory, Revised Edition. 1989

Volume 46 S. GEORGE PHILANDER. El Nino, La Nina, and the Southern Oscillation. 1990

Volume 47 ROBERT A. BROWN. Fluid Mechanics of the Atmosphere. 1991

Volume 48 JAMES R. HOLTON. An Introduction to Dynamic Meteorology, Third Edition. 1992

Volume 49 ALEXANDER A. KAUFMAN. Geophysical Field Theory and Method.
Part A: Gravitational, Electric, and Magnetic Fields. 1992
Part B: Electromagnetic Fields I. 1994
Part C: Electromagnetic Fields II. 1994

Volume 50 SAMUEL S. BUTCHER, GORDON H. ORIANS, ROBERT J. CHARLSON, AND GORDON V. WOLFE. Global Biogeochemical Cycles. 1992

Volume 51 BRIAN EVANS AND TENG-FONG WONG. Fault Mechanics and Transport Properties in Rock. 1992

Volume 52 ROBERT E. HUFFMAN. Atmospheric Ultraviolet Remote Sensing. 1992

Volume 53 ROBERT A. HOUZE, JR. Cloud Dynamics. 1993

Volume 54 PETER V. HOBBS. Aerosol–Cloud–Climate Interactions. 1993

Volume 55 S. J. GIBOWICZ AND A. KIJKO. An Introduction to Mining Seismology. 1993

Volume 56 DENNIS L. HARTMANN. Global Physical Climatology. 1994.

Volume 57 MICHAEL P. RYAN. Magmatic Systems. 1994

Volume 58 THORNE LAY AND TERRY C. WALLACE. Modern Global Seismology. 1995

Volume 59 DANIEL S. WILKS. Statistical Methods in the Atmospheric Sciences. 1995

Volume 60 FREDERIK NEBEKER. Calculating the Weather. In Preparation

*Out of print.

ISBN 0-12-751965-3

90038

9 780127 519654