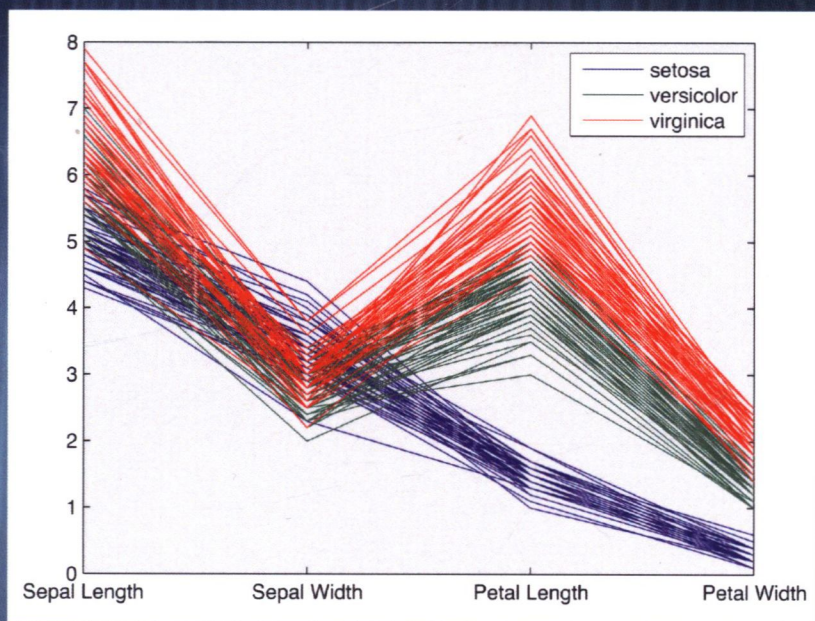


Computer Science and Data Analysis Series

# STATISTICS IN MATLAB®

## *A PRIMER*



**WENDY L. MARTINEZ**  
**MOONJUNG CHO**



**CRC Press**  
Taylor & Francis Group

A CHAPMAN & HALL BOOK



## Computer Science and Data Analysis Series

Fulfilling the need for a practical user's guide, **Statistics in MATLAB: A Primer** provides an accessible introduction to the latest version of MATLAB® and its extensive functionality for statistics. Assuming a basic knowledge of statistics and probability as well as a fundamental understanding of linear algebra concepts, this book:

- Covers capabilities in the main MATLAB package, the Statistics Toolbox, and the student version of MATLAB
- Presents examples of how MATLAB can be used to analyze data
- Offers access to a companion website with data sets and additional examples
- Contains figures and visual aids to assist in application of the software
- Explains how to determine what method should be used for analysis

**Statistics in MATLAB: A Primer** is an ideal reference for undergraduate and graduate students in engineering, mathematics, statistics, economics, biostatistics, and computer science. It is also appropriate for a diverse professional market, making it a valuable addition to the libraries of researchers in statistics, computer science, data mining, machine learning, image analysis, signal processing, and engineering.



**Authors:** Wendy L. Martinez & MoonJung Cho



**CRC Press**

Taylor & Francis Group  
an informa business

[www.crcpress.com](http://www.crcpress.com)





Computer Science and Data Analysis Series

# **STATISTICS IN MATLAB®**

## ***A PRIMER***

**WENDY L. MARTINEZ**

BUREAU OF LABOR STATISTICS  
WASHINGTON, D.C., USA

**MOONJUNG CHO**

BUREAU OF LABOR STATISTICS  
WASHINGTON, D.C., USA



**CRC Press**

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an **informa** business  
**A CHAPMAN & HALL BOOK**

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press  
Taylor & Francis Group  
6000 Broken Sound Parkway NW, Suite 300  
Boca Raton, FL 33487-2742

First issued in hardback 2018

© 2015 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

ISBN 13: 978-1-138-46931-0 (hbk)  
ISBN 13: 978-1-4665-9656-6 (pbk)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at  
<http://www.taylorandfrancis.com>

and the CRC Press Web site at  
<http://www.crcpress.com>

**STATISTICS  
IN MATLAB®**  
*A PRIMER*

Chapman & Hall/CRC

## **Computer Science and Data Analysis Series**

The interface between the computer and statistical sciences is increasing, as each discipline seeks to harness the power and resources of the other. This series aims to foster the integration between the computer sciences and statistical, numerical, and probabilistic methods by publishing a broad range of reference works, textbooks, and handbooks.

### **SERIES EDITORS**

David Blei, Princeton University

David Madigan, Rutgers University

Marina Meila, University of Washington

Fionn Murtagh, Royal Holloway, University of London

Proposals for the series should be sent directly to one of the series editors above, or submitted to:

### **Chapman & Hall/CRC**

Taylor and Francis Group

3 Park Square, Milton Park

Abingdon, OX14 4RN, UK

---

### **Published Titles**

Semisupervised Learning for Computational Linguistics

*Steven Abney*

Visualization and Verbalization of Data

*Jörg Blasius and Michael Greenacre*

Design and Modeling for Computer Experiments

*Kai-Tai Fang, Runze Li, and Agus Sudjianto*

Microarray Image Analysis: An Algorithmic Approach

*Karl Fraser, Zidong Wang, and Xiaohui Liu*

R Programming for Bioinformatics

*Robert Gentleman*

Exploratory Multivariate Analysis by Example Using R

*François Husson, Sébastien Lê, and Jérôme Pagès*

Bayesian Artificial Intelligence, Second Edition

*Kevin B. Korb and Ann E. Nicholson*



## **Published Titles cont.**

Computational Statistics Handbook with MATLAB<sup>®</sup>, Second Edition

*Wendy L. Martinez and Angel R. Martinez*

Exploratory Data Analysis with MATLAB<sup>®</sup>, Second Edition

*Wendy L. Martinez, Angel R. Martinez, and Jeffrey L. Solka*

Statistics in MATLAB<sup>®</sup>: A Primer

*Wendy L. Martinez and MoonJung Cho*

Clustering for Data Mining: A Data Recovery Approach, Second Edition

*Boris Mirkin*

Introduction to Machine Learning and Bioinformatics

*Sushmita Mitra, Sujay Datta, Theodore Perkins, and George Michailidis*

Introduction to Data Technologies

*Paul Murrell*

R Graphics

*Paul Murrell*

Correspondence Analysis and Data Coding with Java and R

*Fionn Murtagh*

Pattern Recognition Algorithms for Data Mining

*Sankar K. Pal and Pabitra Mitra*

Statistical Computing with R

*Maria L. Rizzo*

Statistical Learning and Data Science

*Mireille Gettler Summa, Léon Bottou, Bernard Goldfarb, Fionn Murtagh,*

*Catherine Pardoux, and Myriam Touati*

Foundations of Statistical Algorithms: With References to R Packages

*Claus Weihs, Olaf Mersmann, and Uwe Ligges*





*Wendy dedicates this book to her parents  
who started her on this path:*

*Shirley and Glenn Cukr*

*MoonJung dedicates this book to her children:*

*Catherine and Ted*



---

# Preface

---

The functionality in MATLAB® for statistical data analysis has improved and expanded in the past several years, with major changes made in 2012 (version 8). Additionally, MATLAB is frequently used in academia to teach statistics, engineering, mathematics, and data analysis courses. Thus, we felt that a book that provides an overview or introduction to the extensive functionality available in MATLAB would be useful to a wide audience.

The main MATLAB software includes many basic functions for statistical visualization and data analysis. The MathWorks, Inc. Statistics Toolbox extends these basic capabilities by including additional specialized functions. The Statistics Toolbox can be purchased separately. MathWorks also has a Student Version of MATLAB that includes the Statistics Toolbox, as well as many other toolboxes.

One should have the Statistics Toolbox to get the most from this book. However, we include enough content to help those who do not have the extra MATLAB functionality. We have been careful to note where to find the functions—in the base MATLAB or the Statistics Toolbox. If the reader is ever confused about where the function comes from, type

**which functionname**

at the command line, and the location of the function file will be displayed.

For example, **normpdf** is a function in the Statistics Toolbox. We get the following, when we type **which normpdf**:

**C:\MATLAB2013a\toolbox\stats\stats\normpdf.m**

This is under the directory **~\toolbox\stats**, indicating the function is part of the Statistics Toolbox. A function that is in base MATLAB will have a directory **~\toolbox\matlab**.

It took over a year to write this book and several versions of MATLAB. We started with MATLAB R2013a and finished with MATLAB 2014a. All functions should work with R2013, with a few exceptions. These are found mostly in Chapter 8. We recommend that readers investigate the changes in MATLAB by looking at the release notes that are available in the MATLAB documentation or try to have the latest version installed.

There is a companion website, where the reader can find the data sets, M-files with the code from the book, and additional examples. This website is

**[www.pi-sigma.info](http://www.pi-sigma.info)**

You can also download the files from the book website at CRC Press.

For the most part, we assume the reader has a basic knowledge of statistics and probability. The focus of this book is on how to use the statistics capabilities in MATLAB, not on the theory and use of statistics. However, we do include definitions and formulas in certain areas in order to aid the understanding of the MATLAB functions.

The reader should also know some basic concepts of linear algebra. This includes the definitions of vectors, matrices and operations such as adding vectors, multiplying matrices, taking transposes, etc.

Please note that this book is an *introduction*. It is not meant to cover all aspects of statistical analysis nor all of the functions available for statistics in MATLAB. Readers should always refer to the help files and other documentation provided with MATLAB to get the full story. We provide information on how to access the documentation in the first chapter.

We would like to acknowledge the invaluable help of the reviewers: Tom Lane, John Eltinge, Terrance Savitsky, Eungchun Cho, Ted Cho, and Angel Martinez. Their many helpful comments and suggestions resulted in a better book. Any shortcomings are the sole responsibility of the authors. We greatly appreciate the help and patience of those at CRC Press: David Grubbs, Jessica Vakili, Robin Starkes, and Kevin Craig. Finally, we are indebted to Naomi Fernandes, Paul Pilotte, and Tom Lane at The MathWorks, Inc. for their special assistance with MATLAB.

## Disclaimers

1. Any MATLAB programs and data sets that are included with the book are provided in good faith. The authors, publishers, or distributors do not guarantee their accuracy and are not responsible for the consequences of their use.
2. MATLAB® and Simulink® are trademarks of the MathWorks, Inc. and are used with permission. The MathWorks does not warrant the accuracy of the text or the exercises in this book. This book's use or discussion of MATLAB® and Simulink® software or related products does not constitute endorsement or sponsorship by the MathWorks of a particular pedagogical approach or particular use of the MATLAB® and Simulink® software.
3. The views expressed in this book are those of the authors and do not necessarily represent the views of the United States Department of Labor or its components.

Wendy Martinez  
MoonJung Cho



---

# Table of Contents

---

List of Figures ..... xi

List of Tables ..... xvii

Preface .....xxi

## Chapter 1

### MATLAB® Basics

1.1 Desktop Environment ..... 1

1.2 Getting Help and Other Documentation ..... 4

1.3 Data Import and Export ..... 6

    1.3.1 Data I/O via the Command Line ..... 6

    1.3.2 The Import Wizard ..... 8

    1.3.3 Examples of Data I/O in MATLAB® ..... 8

    1.3.4 Data I/O with the Statistics Toolbox ..... 11

    1.3.5 More Functions for Data I/O ..... 13

1.4 Data in MATLAB® ..... 14

    1.4.1 Data Objects in Base MATLAB® ..... 14

    1.4.2 Accessing Data Elements ..... 20

    1.4.3 Examples of Joining Data Sets ..... 22

    1.4.4 Data Types in the Statistics Toolbox ..... 24

    1.4.5 Object-Oriented Programming ..... 25

1.5 Miscellaneous Topics ..... 27

    1.5.1 File and Workspace Management ..... 27

    1.5.2 Punctuation in MATLAB® ..... 27

    1.5.3 Arithmetic Operators ..... 29

    1.5.4 Functions in MATLAB® ..... 30

1.6 Summary and Further Reading ..... 32

## Chapter 2

### Visualizing Data

2.1 Basic Plot Functions ..... 35

    2.1.1 Plotting 2-D Data ..... 35

    2.1.2 Plotting 3-D Data ..... 39

    2.1.3 Examples ..... 40

2.2 Scatter Plots .....	42
2.2.1 Basic 2-D and 3-D Scatter Plots .....	43
2.2.2 Scatter Plot Matrix .....	44
2.2.3 Examples .....	44
2.3 GUIs for Graphics .....	45
2.3.1 Simple Plot Editing .....	47
2.3.2 Plotting Tools Interface .....	48
2.3.3 <b>Plots</b> Tab .....	49
2.4 Summary and Further Reading .....	51

## Chapter 3

### Descriptive Statistics

3.1 Measures of Location .....	53
3.1.1 Means, Medians, and Modes .....	54
3.1.2 Examples .....	55
3.2 Measures of Dispersion .....	56
3.2.1 Range .....	57
3.2.2 Variance and Standard Deviation .....	58
3.2.3 Covariance and Correlation .....	58
3.2.4 Examples .....	59
3.3 Describing the Distribution .....	62
3.3.1 Quantiles .....	62
3.3.2 Interquartile Range .....	63
3.3.3 Skewness .....	63
3.3.4 Examples .....	65
3.4 Visualizing the Data Distribution .....	66
3.4.1 Histograms .....	66
3.4.2 Probability Plots .....	67
3.4.3 Boxplots .....	68
3.4.4 Examples .....	69
3.5 Summary and Further Reading .....	72

## Chapter 4

### Probability Distributions

4.1 Distributions in MATLAB® .....	77
4.1.1 Continuous Distributions .....	78
4.1.2 Discrete Distributions .....	80
4.1.3 Probability Distribution Objects .....	81
4.1.4 Other Distributions .....	83
4.1.5 Examples of Probability Distributions in MATLAB® .....	87
4.1.6 <b>disttool</b> for Exploring Probability Distributions .....	91
4.2 Parameter Estimation .....	94
4.2.1 Command Line Functions .....	94
4.2.2 Examples of Parameter Estimation .....	95
4.2.3 <b>dfittool</b> for Interactive Fitting .....	100

4.3 Generating Random Numbers .....	105
4.3.1 Generating Random Variables in Base MATLAB® .....	105
4.3.2 Generating Random Variables with the Statistics Toolbox .....	106
4.3.3 Examples of Random Number Generation .....	107
4.3.4 <b>randtool</b> for Generating Random Variables .....	111
4.4 Summary and Further Reading .....	112

## Chapter 5

### Hypothesis Testing

5.1 Basic Concepts .....	115
5.1.1 Hypothesis Testing .....	116
5.1.2 Confidence Intervals .....	118
5.2 Common Hypothesis Tests .....	119
5.2.1 The z-Test and t-Test .....	119
5.2.2 Examples of Hypothesis Tests .....	122
5.3 Confidence Intervals Using Bootstrap Resampling .....	129
5.3.1 The Basic Bootstrap .....	129
5.3.2 Examples .....	130
5.4 Analysis of Variance .....	133
5.4.1 One-Way ANOVA .....	134
5.4.2 ANOVA Example .....	137
5.5 Summary and Further Reading .....	140

## Chapter 6

### Model-Building with Regression Analysis

6.1 Introduction to Linear Models .....	143
6.1.1 Specifying Models .....	144
6.1.2 The Least Squares Approach for Estimation .....	145
6.1.3 Assessing Model Estimates .....	147
6.2 Model-Building Functions in Base MATLAB® .....	148
6.2.1 Fitting Polynomials .....	149
6.2.2 Using the Division Operators .....	153
6.2.3 Ordinary Least Squares .....	155
6.3 Functions in the Statistics Toolbox .....	157
6.3.1 Using <b>regress</b> for Regression Analysis .....	159
6.3.2 Using <b>regstats</b> for Regression Analysis .....	160
6.3.3 The Linear Regression Model Class .....	164
6.3.4 Assessing Model Fit .....	168

6.4 Basic Fitting GUI ..... 176

6.5 Summary and Further Reading ..... 180

**Chapter 7**

**Multivariate Analysis**

7.1 Principal Component Analysis ..... 183

    7.1.1 Functions for PCA in Base MATLAB® ..... 186

    7.1.2 Functions for PCA in the Statistics Toolbox ..... 190

    7.1.3 Biplots ..... 194

7.2 Multidimensional Scaling—MDS ..... 194

    7.2.1 Measuring Distance ..... 196

    7.2.2 Classical MDS ..... 198

    7.2.3 Metric MDS ..... 200

    7.2.4 Nonmetric MDS ..... 203

7.3 Visualization in Higher Dimensions ..... 209

    7.3.1 Scatter Plot Matrix ..... 210

    7.3.2 Parallel Coordinate Plots ..... 215

    7.3.3 Andrews Curves ..... 216

7.4 Summary and Further Reading ..... 219

**Chapter 8**

**Classification and Clustering**

8.1 Supervised Learning or Classification ..... 221

    8.1.1 Bayes Decision Theory ..... 222

    8.1.2 Discriminant Analysis ..... 224

    8.1.3 Naive Bayes Classifiers ..... 228

    8.1.4 Nearest Neighbor Classifier ..... 230

8.2 Unsupervised Learning or Cluster Analysis ..... 233

    8.2.1 Hierarchical Clustering ..... 234

    8.2.2 K-Means Clustering ..... 241

8.3 Summary and Further Reading ..... 249

References ..... 253

Index of MATLAB® Functions ..... 257

Subject Index ..... 261