

Brani Vidakovic

# Statistics for Bioengineering Sciences

With MATLAB and WinBUGS Support

Springer



# Preface

This text is a result of many semesters of teaching introductory statistical courses to engineering students at Duke University and the Georgia Institute of Technology. Through its scope and depth of coverage, the text addresses the needs of the vibrant and rapidly growing engineering fields, bioengineering and biomedical engineering, while implementing software that engineers are familiar with.

There are many good introductory statistics books for engineers on the market, as well as many good introductory biostatistics books. This text is an attempt to put the two together as a single textbook heavily oriented to computation and hands-on approaches. For example, the aspects of disease and device testing, sensitivity, specificity and ROC curves, epidemiological risk theory, survival analysis, and logistic and Poisson regressions are not typical topics for an introductory engineering statistics text. On the other hand, the books in biostatistics are not particularly challenging for the level of computational sophistication that engineering students possess.

The approach enforced in this text avoids the use of mainstream statistical packages in which the procedures are often black-boxed. Rather, the students are expected to code the procedures on their own. The results may not be as flashy as they would be if the specialized packages were used, but the student will go through the process and understand each step of the program. The computational support for this text is the MATLAB<sup>®</sup> programming environment since this software is predominant in the engineering communities. For instance, Georgia Tech has developed a practical introductory course in computing for engineers (CS1371 – Computing for Engineers) that relies on MATLAB. Over 1,000 students take this class per semester as it is a requirement for all engineering students and a prerequisite for many upper-level courses.

In addition to the synergy of engineering and biostatistical approaches, the novelty of this book is in the substantial coverage of Bayesian approaches to statistical inference.




I avoided taking sides on the traditional (classical, frequentist) vs. Bayesian approach; it was my goal to expose students to both approaches. It is undeniable that classical statistics is overwhelmingly used in conducting and reporting inference among practitioners, and that Bayesian statistics is gaining in popularity, acceptance, and usage (FDA, Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, 5 February 2010). Many examples in this text are solved using both the traditional and Bayesian methods, and the results are compared and commented upon.


This diversification is made possible by advances in Bayesian computation and the availability of the free software WinBUGS that provides painless computational support for Bayesian solutions. WinBUGS and MATLAB communicate well due to the free interface software MATBUGS. The book also relies on `stat` toolbox within MATLAB.

The World Wide Web (WWW) facilitates the text. All custom-made MATLAB and WinBUGS programs (compatible with MATLAB 7.12 (2011a) and WinBUGS 1.4.3 or OpenBUGS 3.2.1) as well as data sets used in this book are available on the Web:


<http://springer.bme.gatech.edu/>

To keep the text as lean as possible, solutions and hints to the majority of exercises can be found on the book's Web site. The computer scripts and examples are an integral part of the text, and all MATLAB codes and outputs are shown in *blue typewriter font* while all WinBUGS programs are given in *red-brown typewriter font*. The comments in MATLAB and WinBUGS codes are presented in *green typewriter font*.

The three icons , , and  are used to point to data sets, MATLAB codes, and WinBUGS codes, respectively.

The difficulty of the material in the text necessarily varies. More difficult sections that may be omitted in the basic coverage are denoted by a star, \*. However, it is my experience that advanced undergraduate bioengineering students affiliated with school research labs need and use the “starred” material, such as functional ANOVA, variance stabilizing transforms, and nested experimental designs, to name just a few. Tricky or difficult places are marked with Donald Knut's “bend” .

Each chapter starts with a box titled WHAT IS COVERED IN THIS CHAPTER and ends with chapter exercises, a box called MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER, and chapter references.

The examples are numbered and the end of each example is marked with .

I am aware that this work is not perfect and that many improvements could be made with respect to both exposition and coverage. Thus, I would welcome any criticism and pointers from readers as to how this book could be improved.

**Acknowledgments.** I am indebted to many students and colleagues who commented on various drafts of the book. In particular I am grateful to colleagues from the Department of Biomedical Engineering at the Georgia Institute of Technology and Emory University and their undergraduate and graduate advisees/researchers who contributed with real-life examples and exercises from their research labs.

Colleagues Tom Bylander of the University of Texas at San Antonio, John H. McDonald of the University of Delaware, and Roger W. Johnson of the South Dakota School of Mines & Technology kindly gave permission to use their data and examples. I also acknowledge Mathworks' statistical gurus Peter Perkins and Tom Lane for many useful conversations over the last several years. Several MATLAB codes used in this book come from the MATLAB Central File Exchange forum. In particular, I am grateful to Antonio Truillo-Ortiz and his team (Universidad Autonoma de Baja California) and to Giuseppe Cardillo (Merigen Research) for their excellent contributions.

The book benefited from the input of many diligent students when it was used either as a supplemental reading or later as a draft textbook for a semester-long course at Georgia Tech: BMED2400 Introduction to Bioengineering Statistics. A complete list of students who provided useful comments would be quite long, but the most diligent ones were Erin Hamilton, Kiersten Petersen, David Dreyfus, Jessica Kanter, Radu Reit, Amoreth Gozo, Nader Aboujamous, and Allison Chan.

Springer's team kindly helped along the way. I am grateful to Marc Strauss and Kathryn Schell for their encouragement and support and to Glenn Corey for his knowledgeable copyediting.

Finally, it hardly needs stating that the book would have been considerably less fun to write without the unconditional support of my family.

BRANI VIDAKOVIC  
*School of Biomedical Engineering*  
*Georgia Institute of Technology*  
brani@bme.gatech.edu



# Contents

<b>Preface</b> .....	v
<b>1 Introduction</b> .....	1
Chapter References .....	7
<b>2 The Sample and Its Properties</b> .....	9
2.1 Introduction .....	9
2.2 A MATLAB Session on Univariate Descriptive Statistics .....	10
2.3 Location Measures .....	13
2.4 Variability Measures .....	16
2.5 Displaying Data .....	24
2.6 Multidimensional Samples: Fisher’s Iris Data and Body Fat Data .....	28
2.7 Multivariate Samples and Their Summaries* .....	33
2.8 Visualizing Multivariate Data .....	38
2.9 Observations as Time Series .....	42
2.10 About Data Types .....	44
2.11 Exercises .....	46
Chapter References .....	57
<b>3 Probability, Conditional Probability, and Bayes’ Rule</b> .....	59
3.1 Introduction .....	59
3.2 Events and Probability .....	60
3.3 Odds .....	71
3.4 Venn Diagrams* .....	71
3.5 Counting Principles* .....	74
3.6 Conditional Probability and Independence .....	78
3.6.1 Pairwise and Global Independence .....	82
3.7 Total Probability .....	83
3.8 Bayes’ Rule .....	85
3.9 Bayesian Networks* .....	90

3.10 Exercises .....	96
Chapter References .....	106
<b>4 Sensitivity, Specificity, and Relatives .....</b>	<b>109</b>
4.1 Introduction .....	109
4.2 Notation .....	110
4.2.1 Conditional Probability Notation .....	113
4.3 Combining Two or More Tests .....	115
4.4 ROC Curves .....	118
4.5 Exercises .....	122
Chapter References .....	129
<b>5 Random Variables .....</b>	<b>131</b>
5.1 Introduction .....	131
5.2 Discrete Random Variables .....	133
5.2.1 Jointly Distributed Discrete Random Variables .....	138
5.3 Some Standard Discrete Distributions .....	140
5.3.1 Discrete Uniform Distribution .....	140
5.3.2 Bernoulli and Binomial Distributions .....	141
5.3.3 Hypergeometric Distribution .....	146
5.3.4 Poisson Distribution .....	149
5.3.5 Geometric Distribution .....	151
5.3.6 Negative Binomial Distribution .....	152
5.3.7 Multinomial Distribution .....	155
5.3.8 Quantiles .....	156
5.4 Continuous Random Variables .....	157
5.4.1 Joint Distribution of Two Continuous Random Variables .....	158
5.5 Some Standard Continuous Distributions .....	161
5.5.1 Uniform Distribution .....	161
5.5.2 Exponential Distribution .....	162
5.5.3 Normal Distribution .....	164
5.5.4 Gamma Distribution .....	165
5.5.5 Inverse Gamma Distribution .....	166
5.5.6 Beta Distribution .....	167
5.5.7 Double Exponential Distribution .....	168
5.5.8 Logistic Distribution .....	169
5.5.9 Weibull Distribution .....	170
5.5.10 Pareto Distribution .....	171
5.5.11 Dirichlet Distribution .....	172
5.6 Random Numbers and Probability Tables .....	173
5.7 Transformations of Random Variables* .....	174
5.8 Mixtures* .....	177
5.9 Markov Chains* .....	178
5.10 Exercises .....	180
Chapter References .....	189



<b>6</b>	<b>Normal Distribution</b>	191
6.1	Introduction	191
6.2	Normal Distribution	192
6.2.1	Sigma Rules	197
6.2.2	Bivariate Normal Distribution*	197
6.3	Examples with a Normal Distribution	199
6.4	Combining Normal Random Variables	202
6.5	Central Limit Theorem	204
6.6	Distributions Related to Normal	208
6.6.1	Chi-square Distribution	209
6.6.2	(Student's) $t$ -Distribution	213
6.6.3	Cauchy Distribution	214
6.6.4	$F$ -Distribution	215
6.6.5	Noncentral $\chi^2$ , $t$ , and $F$ Distributions	216
6.6.6	Lognormal Distribution	218
6.7	Delta Method and Variance Stabilizing Transformations*	219
6.8	Exercises	222
	Chapter References	228
<b>7</b>	<b>Point and Interval Estimators</b>	229
7.1	Introduction	229
7.2	Moment Matching and Maximum Likelihood Estimators	230
7.2.1	Unbiasedness and Consistency of Estimators	238
7.3	Estimation of a Mean, Variance, and Proportion	240
7.3.1	Point Estimation of Mean	240
7.3.2	Point Estimation of Variance	242
7.3.3	Point Estimation of Population Proportion	245
7.4	Confidence Intervals	246
7.4.1	Confidence Intervals for the Normal Mean	247
7.4.2	Confidence Interval for the Normal Variance	249
7.4.3	Confidence Intervals for the Population Proportion	253
7.4.4	Confidence Intervals for Proportions When $X = 0$	257
7.4.5	Designing the Sample Size with Confidence Intervals	258
7.5	Prediction and Tolerance Intervals*	260
7.6	Confidence Intervals for Quantiles*	262
7.7	Confidence Intervals for the Poisson Rate*	263
7.8	Exercises	265
	Chapter References	276
<b>8</b>	<b>Bayesian Approach to Inference</b>	279
8.1	Introduction	279
8.2	Ingredients for Bayesian Inference	282
8.3	Conjugate Priors	287
8.4	Point Estimation	288
8.5	Prior Elicitation	290

8.6	Bayesian Computation and Use of WinBUGS	293
8.6.1	Zero Tricks in WinBUGS	296
8.7	Bayesian Interval Estimation: Credible Sets	298
8.8	Learning by Bayes' Theorem	301
8.9	Bayesian Prediction	302
8.10	Consensus Means*	305
8.11	Exercises	308
	Chapter References	314
<b>9</b>	<b>Testing Statistical Hypotheses</b>	<b>317</b>
9.1	Introduction	317
9.2	Classical Testing Problem	319
9.2.1	Choice of Null Hypothesis	319
9.2.2	Test Statistic, Rejection Regions, Decisions, and Errors in Testing	320
9.2.3	Power of the Test	322
9.2.4	Fisherian Approach: $p$ -Values	323
9.3	Bayesian Approach to Testing	324
9.3.1	Criticism and Calibration of $p$ -Values*	327
9.4	Testing the Normal Mean	329
9.4.1	$z$ -Test	329
9.4.2	Power Analysis of a $z$ -Test	330
9.4.3	Testing a Normal Mean When the Variance Is Not Known: $t$ -Test	331
9.4.4	Power Analysis of $t$ -Test	335
9.5	Testing the Normal Variances	336
9.6	Testing the Proportion	338
9.7	Multiplicity in Testing, Bonferroni Correction, and False Discovery Rate	341
9.8	Exercises	344
	Chapter References	353
<b>10</b>	<b>Two Samples</b>	<b>355</b>
10.1	Introduction	355
10.2	Means and Variances in Two Independent Normal Populations	356
10.2.1	Confidence Interval for the Difference of Means	361
10.2.2	Power Analysis for Testing Two Means	361
10.2.3	More Complex Two-Sample Designs	363
10.2.4	Bayesian Test of Two Normal Means	365
10.3	Testing the Equality of Normal Means When Samples Are Paired	367
10.3.1	Sample Size in Paired $t$ -Test	373
10.4	Two Variances	373
10.5	Comparing Two Proportions	378
10.5.1	The Sample Size	379

10.6 Risks: Differences, Ratios, and Odds Ratios .....	380
10.6.1 Risk Differences .....	381
10.6.2 Risk Ratio .....	382
10.6.3 Odds Ratios .....	383
10.7 Two Poisson Rates* .....	387
10.8 Equivalence Tests* .....	389
10.9 Exercises .....	393
Chapter References .....	406
<b>11 ANOVA and Elements of Experimental Design .....</b>	<b>409</b>
11.1 Introduction .....	409
11.2 One-Way ANOVA .....	410
11.2.1 ANOVA Table and Rationale for F-Test .....	412
11.2.2 Testing Assumption of Equal Population Variances .....	415
11.2.3 The Null Hypothesis Is Rejected. What Next? .....	416
11.2.4 Bayesian Solution .....	421
11.2.5 Fixed- and Random-Effect ANOVA .....	423
11.3 Two-Way ANOVA and Factorial Designs .....	424
11.4 Blocking .....	430
11.5 Repeated Measures Design .....	431
11.5.1 Sphericity Tests .....	435
11.6 Nested Designs* .....	436
11.7 Power Analysis in ANOVA .....	438
11.8 Functional ANOVA* .....	443
11.9 Analysis of Means (ANOM)* .....	446
11.10 Gauge R&R ANOVA* .....	448
11.11 Testing Equality of Several Proportions .....	454
11.12 Testing the Equality of Several Poisson Means* .....	455
11.13 Exercises .....	457
Chapter References .....	475
<b>12 Distribution-Free Tests .....</b>	<b>477</b>
12.1 Introduction .....	477
12.2 Sign Test .....	478
12.3 Ranks .....	481
12.4 Wilcoxon Signed-Rank Test .....	483
12.5 Wilcoxon Sum Rank Test and Wilcoxon–Mann–Whitney Test .....	486
12.6 Kruskal–Wallis Test .....	490
12.7 Friedman’s Test .....	492
12.8 Walsh Nonparametric Test for Outliers* .....	495
12.9 Exercises .....	496
Chapter References .....	500

<b>13</b>	<b>Goodness-of-Fit Tests</b>	503
13.1	Introduction	503
13.2	Quantile–Quantile Plots	504
13.3	Pearson’s Chi-Square Test	508
13.4	Kolmogorov–Smirnov Tests	515
13.4.1	Kolmogorov’s Test	515
13.4.2	Smirnov’s Test to Compare Two Distributions	517
13.5	Moran’s Test*	520
13.6	Departures from Normality	521
13.7	Exercises	523
	Chapter References	529
<b>14</b>	<b>Models for Tables</b>	531
14.1	Introduction	531
14.2	Contingency Tables: Testing for Independence	532
14.2.1	Measuring Association in Contingency Tables	537
14.2.2	Cohen’s Kappa	540
14.3	Three-Way Tables	543
14.4	Fisher’s Exact Test	546
14.5	Multiple Tables: Mantel–Haenszel Test	548
14.5.1	Testing Conditional Independence or Homogeneity	549
14.5.2	Conditional Odds Ratio	551
14.6	Paired Tables: McNemar’s Test	552
14.6.1	Risk Differences	553
14.6.2	Risk Ratios	554
14.6.3	Odds Ratios	554
14.6.4	Stuart–Maxwell Test*	559
14.7	Exercises	561
	Chapter References	569
<b>15</b>	<b>Correlation</b>	571
15.1	Introduction	571
15.2	The Pearson Coefficient of Correlation	572
15.2.1	Inference About $\rho$	574
15.2.2	Bayesian Inference for Correlation Coefficients	585
15.3	Spearman’s Coefficient of Correlation	586
15.4	Kendall’s Tau	589
15.5	<i>Cum hoc ergo propter hoc</i>	591
15.6	Exercises	592
	Chapter References	596

<b>16</b>	<b>Regression</b> .....	599
	16.1 Introduction .....	599
	16.2 Simple Linear Regression .....	600
	16.2.1 Testing Hypotheses in Linear Regression .....	608
	16.3 Testing the Equality of Two Slopes* .....	616
	16.4 Multivariable Regression .....	619
	16.4.1 Matrix Notation .....	620
	16.4.2 Residual Analysis, Influential Observations, Multicollinearity, and Variable Selection* .....	625
	16.5 Sample Size in Regression .....	634
	16.6 Linear Regression That Is Nonlinear in Predictors .....	635
	16.7 Errors-In-Variables Linear Regression* .....	637
	16.8 Analysis of Covariance .....	638
	16.9 Exercises .....	644
	Chapter References .....	656
<b>17</b>	<b>Regression for Binary and Count Data</b> .....	657
	17.1 Introduction .....	657
	17.2 Logistic Regression .....	658
	17.2.1 Fitting Logistic Regression .....	659
	17.2.2 Assessing the Logistic Regression Fit .....	664
	17.2.3 Probit and Complementary Log-Log Links .....	674
	17.3 Poisson Regression .....	678
	17.4 Log-linear Models .....	684
	17.5 Exercises .....	688
	Chapter References .....	699
<b>18</b>	<b>Inference for Censored Data and Survival Analysis</b> .....	701
	18.1 Introduction .....	701
	18.2 Definitions .....	702
	18.3 Inference with Censored Observations .....	704
	18.3.1 Parametric Approach .....	704
	18.3.2 Nonparametric Approach: Kaplan–Meier Estimator .....	706
	18.3.3 Comparing Survival Curves .....	712
	18.4 The Cox Proportional Hazards Model .....	714
	18.5 Bayesian Approach .....	718
	18.5.1 Survival Analysis in WinBUGS .....	720
	18.6 Exercises .....	726
	Chapter References .....	730
<b>19</b>	<b>Bayesian Inference Using Gibbs Sampling – BUGS Project</b> ...	733
	19.1 Introduction .....	733
	19.2 Step-by-Step Session .....	734
	19.3 Built-in Functions and Common Distributions in WinBUGS .....	739
	19.4 MATBUGS: A MATLAB Interface to WinBUGS .....	740

19.5 Exercises .....	744
Chapter References .....	745
<b>Index</b> .....	<b>747</b>

# Chapter 1

## Introduction

*Many people were at first surprised at my using the new words “Statistics” and “Statistical,” as it was supposed that some term in our own language might have expressed the same meaning. But in the course of a very extensive tour through the northern parts of Europe, which I happened to take in 1786, I found that in Germany they were engaged in a species of political inquiry to which they had given the name of “Statistics”. . . . I resolved on adopting it, and I hope that it is now completely naturalised and incorporated with our language.*

– Sinclair, 1791; Vol XX

### WHAT IS COVERED IN THIS CHAPTER

- What is the subject of statistics?
- Population, sample, data
- Appetizer examples



The problems confronting health professionals today often involve fundamental aspects of device and system analysis, and their design and application, and as such are of extreme importance to engineers and scientists.

Because many aspects of engineering and scientific practice involve non-deterministic outcomes, understanding and knowledge of statistics is important to any engineer and scientist. Statistics is a *guide to the unknown*. It is a science that deals with designing experimental protocols, collecting, summarizing, and presenting data, and, most importantly, making inferences and