

---

SPRINGER TEXTS IN STATISTICS

---

# All of Statistics

A Concise Course  
in Statistical  
Inference

Larry Wasserman

 Springer

---

# *Springer Texts in Statistics*

*Advisors:*

George Casella   Stephen Fienberg   Ingram Olkin

## Springer Texts in Statistics

---

- Alfred*: Elements of Statistics for the Life and Social Sciences  
*Berger*: An Introduction to Probability and Stochastic Processes  
*Bilodeau and Brenner*: Theory of Multivariate Statistics  
*Blom*: Probability and Statistics: Theory and Applications  
*Brockwell and Davis*: Introduction to Times Series and Forecasting, Second Edition  
*Chow and Teicher*: Probability Theory: Independence, Interchangeability, Martingales, Third Edition  
*Christensen*: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization, Second Edition  
*Christensen*: Log-Linear Models and Logistic Regression, Second Edition  
*Christensen*: Plane Answers to Complex Questions: The Theory of Linear Models, Third Edition  
*Creighton*: A First Course in Probability Models and Statistical Inference  
*Davis*: Statistical Methods for the Analysis of Repeated Measurements  
*Dean and Voss*: Design and Analysis of Experiments  
*du Toit, Steyn, and Stumpf*: Graphical Exploratory Data Analysis  
*Durrett*: Essentials of Stochastic Processes  
*Edwards*: Introduction to Graphical Modelling, Second Edition  
*Finkelstein and Levin*: Statistics for Lawyers  
*Flury*: A First Course in Multivariate Statistics  
*Jobson*: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design  
*Jobson*: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods  
*Kalbfleisch*: Probability and Statistical Inference, Volume I: Probability, Second Edition  
*Kalbfleisch*: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition  
*Karr*: Probability  
*Keyfitz*: Applied Mathematical Demography, Second Edition  
*Kiefer*: Introduction to Statistical Inference  
*Kokoska and Nevson*: Statistical Tables and Formulae  
*Kulkarni*: Modeling, Analysis, Design, and Control of Stochastic Systems  
*Lange*: Applied Probability  
*Lehmann*: Elements of Large-Sample Theory  
*Lehmann*: Testing Statistical Hypotheses, Second Edition  
*Lehmann and Casella*: Theory of Point Estimation, Second Edition  
*Lindman*: Analysis of Variance in Experimental Design  
*Lindsey*: Applying Generalized Linear Models

(continued after index)

---

Larry Wasserman

# All of Statistics

A Concise Course in Statistical Inference

With 95 Figures



Springer

Larry Wasserman  
Department of Statistics  
Carnegie Mellon University  
Baker Hall 278A  
Pittsburgh, PA 15213-3890  
USA  
lwarry@stat.cmu.edu

*Editorial Board*

George Casella  
Department of Statistics  
University of Florida  
Gainesville, FL 32611-8545  
USA

Stephen Fienberg  
Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890  
USA

Ingram Olkin  
Department of Statistics  
Stanford University  
Stanford, CA 94305  
USA

Library of Congress Cataloging-in-Publication Data

Wasserman, Larry A. (Larry Allen), 1959–

All of statistics: a concise course in statistical inference / Larry A. Wasserman.

p. cm. — (Springer texts in statistics)

Includes bibliographical references and index.

1. Mathematical statistics. I. Title. II. Series.

QA276 .L2 W37 2009

519.5—dc22

2009090309

ISBN 978-1-4419-2322-6 ISBN 978-1-397-21756-9 (eBook)

DOI 10.1007/978-1-397-21756-9

© 2004 Springer Science+Business Media New York

Originally published by Springer Science+Business Media, Inc. © 2004

Softcover reprint of the hardcover 1st edition 2004

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC), except for brief excerpts in connection with reviews in scholarly journals.

Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 (Covered second printing, 2005)

springer.com

---

*To Isa*

---

# Preface

Taken literally, the title “All of Statistics” is an exaggeration. But in spirit, the title is apt, as the book does cover a much broader range of topics than a typical introductory book on mathematical statistics.

This book is for people who want to learn probability and statistics quickly. It is suitable for graduate or advanced undergraduate students in computer science, mathematics, statistics, and related disciplines. The book includes modern topics like nonparametric curve estimation, bootstrapping, and classification, topics that are usually relegated to follow-up courses. The reader is presumed to know calculus and a little linear algebra. No previous knowledge of probability and statistics is required.

**Statistics, data mining, and machine learning** are all concerned with collecting and analyzing data. For some time, statistics research was conducted in statistics departments while data mining and machine learning research was conducted in computer science departments. Statisticians thought that computer scientists were reinventing the wheel. Computer scientists thought that statistical theory didn’t apply to their problems.

Things are changing. Statisticians now recognize that computer scientists are making novel contributions while computer scientists now recognize the generality of statistical theory and methodology. Clever data mining algorithms are more scalable than statisticians ever thought possible. *Journal* statistical theory is more pervasive than computer scientists had realized.

Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector

machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid.

But where can students learn basic probability and statistics quickly? Nowhere. At least, that was my conclusion when my computer science colleagues kept asking me: "Where can I send my students to get a good understanding of modern statistics quickly?" The typical mathematical statistics course spends too much time on tedious and uninspiring topics (counting methods, two dimensional integrals, etc.) at the expense of covering modern concepts (bootstrapping, curve estimation, graphical models, etc.). So I set out to redesign our undergraduate honors course on probability and mathematical statistics. This book arose from that course. Here is a summary of the main features of this book.

1. The book is suitable for graduate students in computer science and honors undergraduates in math, statistics, and computer science. It is also useful for students beginning graduate work in statistics who need to fill in their background on mathematical statistics.
2. I cover advanced topics that are traditionally not taught in a first course. For example, nonparametric regression, bootstrapping, density estimation, and graphical models.
3. I have omitted topics in probability that do not play a central role in statistical inference. For example, counting methods are virtually absent.
4. Whenever possible, I avoid tedious calculations in favor of emphasizing concepts.
5. I cover nonparametric inference before parametric inference.
6. I abandon the usual "First Term = Probability" and "Second Term = Statistics" approach. Some students only take the first half and it would be a crime if they did not see any statistical theory. Furthermore, probability is more engaging when students can see it put to work in the context of statistics. An exception is the topic of stochastic processes which is included in the later material.
7. The course moves very quickly and covers much material. My colleagues joke that I cover all of statistics in this course and hence the title. The course is demanding but I have worked hard to make the material as intuitive as possible so that the material is very understandable despite the fast pace.
8. Rigor and clarity are not synonymous. I have tried to strike a good balance. To avoid getting bogged down in uninteresting technical details, many results are stated without proof. The bibliographic references at the end of each chapter point the student to appropriate sources.



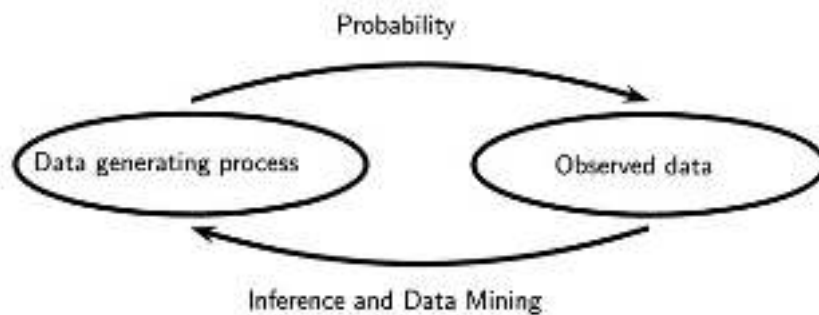


FIGURE 1. Probability and inference.

9. On my website are files with R code which students can use for doing all the computing. The website is:

<http://www.stat.cmu.edu/~larry/all-of-statistics>

However, the book is not tied to R and any computing language can be used.

Part I of the text is concerned with probability theory, the formal language of uncertainty which is the basis of statistical inference. The basic problem that we study in probability is:

Given a data generating process, what are the properties of the outcomes?

Part II is about statistical inference and its close cousins, data mining and machine learning. The basic problem of statistical inference is the inverse of probability:

Given the outcomes, what can we say about the process that generated the data?

These ideas are illustrated in Figure 1. Prediction, classification, clustering, and estimation are all special cases of statistical inference. Data analysis, machine learning and data mining are various names given to the practice of statistical inference, depending on the context.

Part III applies the ideas from Part II to specific problems such as regression, graphical models, causation, density estimation, smoothing, classification, and simulation. Part III contains one more chapter on probability that covers stochastic processes including Markov chains.

I have drawn on other books in many places. Most chapters contain a section called Bibliographic Remarks which serves both to acknowledge my debt to other authors and to point readers to other useful references. I would especially like to mention the books by DeGroot and Schervish (2002) and Grimmett and Stirzaker (1982) from which I adapted many examples and exercises.

As one develops a book over several years it is easy to lose track of where presentation ideas and, especially, homework problems originated. Some I made up. Some I remembered from my education. Some I borrowed from other books. I hope I do not offend anyone if I have used a problem from their book and failed to give proper credit. As my colleague Mark Schervish wrote in his book (Schervish (1995)),

“... the problems at the ends of each chapter have come from many sources. ... These problems, in turn, came from various sources unknown to me ... If I have used a problem without giving proper credit, please take it as a compliment.”

I am indebted to many people without whose help I could not have written this book. First and foremost, the many students who used earlier versions of this text and provided much feedback. In particular, Liz Prather and Jennifer Bakal read the book carefully. Rob Reeder valiantly read through the entire book in excruciating detail and gave me countless suggestions for improvements. Chris Genovese deserves special mention. He not only provided helpful ideas about intellectual content, but also spent many, many hours writing  $\LaTeX$  code for the book. The best aspects of the book's layout are due to his hard work; any stylistic deficiencies are due to my lack of expertise. David Hand, Sam Roweis, and David Scott read the book very carefully and made numerous suggestions that greatly improved the book. John Lafferty and Peter Spirtes also provided helpful feedback. John Kimmel has been supportive and helpful throughout the writing process. Finally, my wife Isabella Verdinelli has been an invaluable source of love, support, and inspiration.

*Larry Wasserman*  
Pittsburgh, Pennsylvania  
July 2003

## Statistics/Data Mining Dictionary

Statisticians and computer scientists often use different language for the same thing. Here is a dictionary that the reader may want to return to throughout the course.

<u>Statistics</u>	<u>Computer Science</u>	<u>Meaning</u>
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete $Y$ from $X$
clustering	unsupervised learning	putting data in groups
data	learning sample	$(X_1, Y_1), \dots, (X_n, Y_n)$
covariates	feature	the $X_i$ 's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space $\Theta$
confidence interval	—	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update beliefs
frequentist inference	-	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

---

# Contents

## I Probability

<b>1 Probability</b>	<b>3</b>
1.1 Introduction	4
1.2 Sample Spaces and Events	4
1.3 Probability	5
1.4 Probability on Finite Sample Spaces	7
1.5 Independent Events	5
1.6 Conditional Probability	10
1.7 Bayes' Theorem	12
1.8 Biographic Remarks	13
1.9 Appendix	13
1.10 Exercises	13
<b>2 Random Variables</b>	<b>19</b>
2.1 Introduction	19
2.2 Distribution Functions and Probability Functions	20
2.3 Some Important Discrete Random Variables	25
2.4 Some Important Continuous Random Variables	27
2.5 Bivariate Distributions	31
2.6 Marginal Distributions	33
2.7 Independent Random Variables	34
2.8 Conditional Distributions	36

2.9	Multivariate Distributions and $n$ Samples . . . . .	38
2.10	Two Important Multivariate Distributions . . . . .	39
2.11	Transformations of Random Variables . . . . .	41
2.12	Transformations of Several Random Variables . . . . .	42
2.13	Appendix . . . . .	43
2.14	Exercises . . . . .	43
<b>3</b>	<b>Expectation</b> . . . . .	<b>47</b>
3.1	Expectation of a Random Variable . . . . .	47
3.2	Properties of Expectations . . . . .	50
3.3	Variance and Covariance . . . . .	50
3.4	Expectation and Variance of Important Random Variables . . . . .	52
3.5	Conditional Expectation . . . . .	54
3.6	Moment Generating Functions . . . . .	56
3.7	Appendix . . . . .	58
3.8	Exercises . . . . .	58
<b>4</b>	<b>Inequalities</b> . . . . .	<b>63</b>
4.1	Probability Inequalities . . . . .	63
4.2	Inequalities For Expectations . . . . .	66
4.3	Bibliographic Remarks . . . . .	66
4.4	Appendix . . . . .	67
4.5	Exercises . . . . .	68
<b>5</b>	<b>Convergence of Random Variables</b> . . . . .	<b>71</b>
5.1	Introduction . . . . .	71
5.2	Types of Convergence . . . . .	72
5.3	The Law of Large Numbers . . . . .	76
5.4	The Central Limit Theorem . . . . .	77
5.5	The Delta Method . . . . .	79
5.6	Bibliographic Remarks . . . . .	80
5.7	Appendix . . . . .	81
	5.7.1 Almost Sure and $L_1$ Convergence . . . . .	81
	5.7.2 Proof of the Central Limit Theorem . . . . .	81
5.8	Exercises . . . . .	82
<b>II</b>	<b>Statistical Inference</b> . . . . .	
<b>6</b>	<b>Models, Statistical Inference and Learning</b> . . . . .	<b>87</b>
6.1	Introduction . . . . .	87
6.2	Parametric and Nonparametric Models . . . . .	87
6.3	Fundamental Concepts in Inference . . . . .	90
	6.3.1 Point Estimation . . . . .	90
	6.3.2 Confidence Sets . . . . .	91

6.3.3	Hypothesis Testing . . . . .	94
6.4	Bibliographic Remarks . . . . .	95
6.5	Appendix . . . . .	95
6.6	Exercises . . . . .	95
<b>7</b>	<b>Estimating the CDF and Statistical Functionals</b>	<b>97</b>
7.1	The Empirical Distribution Function . . . . .	97
7.2	Statistical Functionals . . . . .	99
7.3	Bibliographic Remarks . . . . .	104
7.4	Exercises . . . . .	104
<b>8</b>	<b>The Bootstrap</b>	<b>107</b>
8.1	Simulation . . . . .	108
8.2	Bootstrap Variance Estimation . . . . .	108
8.3	Bootstrap Confidence Intervals . . . . .	110
8.4	Bibliographic Remarks . . . . .	115
8.5	Appendix . . . . .	115
8.5.1	The Jackknife . . . . .	115
8.5.2	Justification For The Percentile Interval . . . . .	116
8.6	Exercises . . . . .	116
<b>9</b>	<b>Parametric Inference</b>	<b>119</b>
9.1	Parameter of Interest . . . . .	120
9.2	The Method of Moments . . . . .	120
9.3	Maximum Likelihood . . . . .	122
9.4	Properties of Maximum Likelihood Estimators . . . . .	124
9.5	Consistency of Maximum Likelihood Estimators . . . . .	126
9.6	Equivariance of the MLE . . . . .	127
9.7	Asymptotic Normality . . . . .	128
9.8	Optimality . . . . .	130
9.9	The Delta Method . . . . .	131
9.10	Multiparameter Models . . . . .	133
9.11	The Parametric Bootstrap . . . . .	134
9.12	Checking Assumptions . . . . .	135
9.13	Appendix . . . . .	135
9.13.1	Proofs . . . . .	135
9.13.2	Sufficiency . . . . .	137
9.13.3	Exponential Families . . . . .	140
9.13.4	Computing Maximum Likelihood Estimates . . . . .	142
9.14	Exercises . . . . .	146
<b>10</b>	<b>Hypothesis Testing and p-values</b>	<b>149</b>
10.1	The Wald Test . . . . .	152
10.2	p-values . . . . .	156
10.3	The $\chi^2$ Distribution . . . . .	159

10.4	Pearson's $\chi^2$ Test for Multinomial Data . . . . .	160
10.5	The Permutation Test . . . . .	161
10.6	The Likelihood Ratio Test . . . . .	161
10.7	Multiple Testing . . . . .	165
10.8	Goodness-of-fit Tests . . . . .	168
10.9	Bibliographic Remarks . . . . .	169
10.10	Appendix . . . . .	170
10.10.1	The Neyman-Pearson Lemma . . . . .	170
10.10.2	The $F$ -test . . . . .	170
10.11	Exercises . . . . .	170
<b>11</b>	<b>Bayesian Inference</b> . . . . .	<b>175</b>
11.1	The Bayesian Philosophy . . . . .	175
11.2	The Bayesian Method . . . . .	176
11.3	Functions of Parameters . . . . .	180
11.4	Simulation . . . . .	180
11.5	Large Sample Properties of Bayes' Procedures . . . . .	181
11.6	Flat Priors, Improper Priors, and "Noninformative" Priors . . . . .	181
11.7	Multi-parameter Problems . . . . .	183
11.8	Bayesian Testing . . . . .	184
11.9	Strengths and Weaknesses of Bayesian Inference . . . . .	185
11.10	Bibliographic Remarks . . . . .	189
11.11	Appendix . . . . .	190
11.12	Exercises . . . . .	190
<b>12</b>	<b>Statistical Decision Theory</b> . . . . .	<b>193</b>
12.1	Preliminaries . . . . .	193
12.2	Comparing Risk Functions . . . . .	194
12.3	Bayes Estimates . . . . .	197
12.4	Minimax Rules . . . . .	198
12.5	Maximum Likelihood, Minimax, and Bayes . . . . .	201
12.6	Admissibility . . . . .	202
12.7	Stein's Paradox . . . . .	203
12.8	Bibliographic Remarks . . . . .	204
12.9	Exercises . . . . .	204
 <b>III Statistical Models and Methods</b>		
<b>13</b>	<b>Linear and Logistic Regression</b> . . . . .	<b>209</b>
13.1	Simple Linear Regression . . . . .	209
13.2	Least Squares and Maximum Likelihood . . . . .	212
13.3	Properties of the Least Squares Estimators . . . . .	214
13.4	Prediction . . . . .	215
13.5	Multiple Regression . . . . .	216

13.6	Model Selection . . . . .	218
13.7	Logistic Regression . . . . .	218
13.8	Bibliographic Remarks . . . . .	215
13.9	Appendix . . . . .	225
13.10	Exercises . . . . .	226
<b>14</b>	<b>Multivariate Models</b>	<b>231</b>
14.1	Random Vectors . . . . .	232
14.2	Estimating the Covariance . . . . .	233
14.3	Multivariate Normal . . . . .	234
14.4	Multinomial . . . . .	235
14.5	Bibliographic Remarks . . . . .	237
14.6	Appendix . . . . .	237
14.7	Exercises . . . . .	235
<b>15</b>	<b>Inference About Independence</b>	<b>239</b>
15.1	Two Binary Variables . . . . .	239
15.2	Two Discrete Variables . . . . .	243
15.3	Two Continuous Variables . . . . .	241
15.4	One Continuous Variable and One Discrete . . . . .	244
15.5	Appendix . . . . .	245
15.6	Exercises . . . . .	248
<b>16</b>	<b>Causal Inference</b>	<b>251</b>
16.1	The Counterfactual Model . . . . .	251
16.2	Beyond Binary Treatments . . . . .	253
16.3	Observational Studies and Confounding . . . . .	257
16.4	Simpson's Paradox . . . . .	259
16.5	Bibliographic Remarks . . . . .	261
16.6	Exercises . . . . .	261
<b>17</b>	<b>Directed Graphs and Conditional Independence</b>	<b>263</b>
17.1	Introduction . . . . .	263
17.2	Conditional Independence . . . . .	264
17.3	DAGs . . . . .	264
17.4	Probability and DAGs . . . . .	266
17.5	More Independence Relations . . . . .	267
17.6	Estimation for DAGs . . . . .	272
17.7	Bibliographic Remarks . . . . .	272
17.8	Appendix . . . . .	272
17.9	Exercises . . . . .	276
<b>18</b>	<b>Undirected Graphs</b>	<b>281</b>
18.1	Undirected Graphs . . . . .	281
18.2	Probability and Graphs . . . . .	289



18.3	Cliques and Potentials . . . . .	284
18.4	Fitting Graphs to Data . . . . .	286
18.5	Bibliographic Remarks . . . . .	289
18.6	Exercises . . . . .	289
<b>19</b>	<b>Log-Linear Models</b> . . . . .	<b>291</b>
19.1	The Log-Linear Model . . . . .	291
19.2	Graphical Log-Linear Models . . . . .	294
19.3	Hierarchical Log-Linear Models . . . . .	296
19.4	Model Generators . . . . .	297
19.5	Fitting Log-Linear Models to Data . . . . .	298
19.6	Bibliographic Remarks . . . . .	300
19.7	Exercises . . . . .	301
<b>20</b>	<b>Nonparametric Curve Estimation</b> . . . . .	<b>303</b>
20.1	The Bias-Variance Tradeoff . . . . .	304
20.2	Histograms . . . . .	306
20.3	Kernel Density Estimation . . . . .	313
20.4	Nonparametric Regression . . . . .	319
20.5	Appendix . . . . .	324
20.6	Bibliographic Remarks . . . . .	325
20.7	Exercises . . . . .	325
<b>21</b>	<b>Smoothing Using Orthogonal Functions</b> . . . . .	<b>327</b>
21.1	Orthogonal Functions and $L_2$ Spaces . . . . .	327
21.2	Density Estimation . . . . .	331
21.3	Regression . . . . .	335
21.4	Wavelets . . . . .	340
21.5	Appendix . . . . .	345
21.6	Bibliographic Remarks . . . . .	346
21.7	Exercises . . . . .	346
<b>22</b>	<b>Classification</b> . . . . .	<b>349</b>
22.1	Introduction . . . . .	349
22.2	Error Rates and the Bayes Classifier . . . . .	350
22.3	Gaussian and Linear Classifiers . . . . .	353
22.4	Linear Regression and Logistic Regression . . . . .	356
22.5	Relationship Between Logistic Regression and LDA . . . . .	358
22.6	Density Estimation and Naive Bayes . . . . .	359
22.7	Trees . . . . .	360
22.8	Assessing Error Rates and Choosing a Good Classifier . . . . .	362
22.9	Support Vector Machines . . . . .	368
22.10	Kernelization . . . . .	371
22.11	Other Classifiers . . . . .	375
22.12	Bibliographic Remarks . . . . .	377

22.13 Exercises . . . . .	377
<b>23 Probability Redux: Stochastic Processes</b>	<b>387</b>
23.1 Introduction . . . . .	387
23.2 Markov Chains . . . . .	388
23.3 Poisson Processes . . . . .	394
23.4 Bibliographic Remarks . . . . .	397
23.5 Exercises . . . . .	398
<b>24 Simulation Methods</b>	<b>403</b>
24.1 Bayesian Inference Revisited . . . . .	403
24.2 Basic Monte Carlo Integration . . . . .	404
24.3 Importance Sampling . . . . .	408
24.4 MCMC Part I: The Metropolis Hastings Algorithm . . . . .	417
24.5 MCMC Part II: Different Flavors . . . . .	418
24.6 Bibliographic Remarks . . . . .	420
24.7 Exercises . . . . .	420
<b>Index</b>	<b>434</b>

---

Part I

**Probability**

---

# 1

## Probability

### 1.1 Introduction

Probability is a mathematical language for quantifying uncertainty. In this Chapter we introduce the basic concepts underlying probability theory. We begin with the **sample space**, which is the set of possible outcomes.

### 1.2 Sample Spaces and Events

The **sample space**  $\Omega$  is the set of possible outcomes of an experiment. Points  $\omega$  in  $\Omega$  are called **sample outcomes**, **realizations**, or **elements**. Subsets of  $\Omega$  are called **Events**.

**1.1 Example.** If we toss a coin twice then  $\Omega = \{HH, HT, TH, TT\}$ . The event that the first toss is heads is  $A = \{HH, HT\}$ . ■

**1.2 Example.** Let  $\omega$  be the outcome of a measurement of some physical quantity, for example, temperature. Then  $\Omega = \mathbb{R} = (-\infty, \infty)$ . One could argue that taking  $\Omega = \mathbb{R}$  is not accurate since temperature has a lower bound. But there is usually no harm in taking the sample space to be larger than needed. The event that the measurement is larger than 10 but less than or equal to 23 is  $A = (10, 23]$ . ■

**1.3 Example.** If we toss a coin forever, then the sample space is the infinite set

$$\Omega = \left\{ \omega = (\omega_1, \omega_2, \omega_3, \dots) : \omega_i \in \{H, T\} \right\}.$$

Let  $E$  be the event that the first Head appears on the third toss. Then

$$E = \left\{ (\omega_1, \omega_2, \omega_3, \dots) : \omega_1 = T, \omega_2 = T, \omega_3 = H, \omega_i \in \{H, T\} \text{ for } i > 3 \right\}. \quad \blacksquare$$

Given an event  $A$ , let  $A^c = \{\omega \in \Omega : \omega \notin A\}$  denote the complement of  $A$ . Informally,  $A^c$  can be read as “not  $A$ .” The complement of  $\Omega$  is the empty set  $\emptyset$ . The union of events  $A$  and  $B$  is defined

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{both}\}$$

which can be thought of as “ $A$  or  $B$ .” If  $A_1, A_2, \dots$  is a sequence of sets then

$$\bigcup_{i=1}^{\infty} A_i = \left\{ \omega \in \Omega : \omega \in A_i \text{ for at least one } i \right\}.$$

The intersection of  $A$  and  $B$  is

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$

read “ $A$  and  $B$ .” Sometimes we write  $A \cap B$  as  $AB$  or  $(A, B)$ . If  $A_1, A_2, \dots$  is a sequence of sets then

$$\bigcap_{i=1}^{\infty} A_i = \left\{ \omega \in \Omega : \omega \in A_i \text{ for all } i \right\}.$$

The set difference is defined by  $A - B = \{\omega : \omega \in A, \omega \notin B\}$ . If every element of  $A$  is also contained in  $B$  we write  $A \subset B$  or, equivalently,  $B \supset A$ . If  $A$  is a finite set, let  $|A|$  denote the number of elements in  $A$ . See the following table for a summary.

	Summary of terminology
$\Omega$	sample space
$\omega$	outcome (point, or element)
$A$	event (subset of $\Omega$ )
$A^c$	complement of $A$ (not $A$ )
$A \cup B$	union ( $A$ or $B$ )
$A \cap B$ or $AB$	intersection ( $A$ and $B$ )
$A - B$	set difference ( $\omega$ in $A$ but not in $B$ )
$A \subset B$	set inclusion
$\emptyset$	null event (always false)
$\Omega$	true event (always true)

We say that  $A_1, A_2, \dots$  are **disjoint** or are **mutually exclusive** if  $A_i \cap A_j = \emptyset$  whenever  $i \neq j$ . For example,  $A_1 = [0, 1), A_2 = [1, 2), A_3 = [2, 3), \dots$  are disjoint. A **partition** of  $\Omega$  is a sequence of disjoint sets  $A_1, A_2, \dots$  such that  $\bigcup_{i=1}^{\infty} A_i = \Omega$ . Given an event  $A$ , define the **indicator function** of  $A$  by

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

A sequence of sets  $A_1, A_2, \dots$  is **monotone increasing** if  $A_1 \subset A_2 \subset \dots$  and we define  $\lim_{n \rightarrow \infty} A_n = \bigcup_{i=1}^{\infty} A_i$ . A sequence of sets  $A_1, A_2, \dots$  is **monotone decreasing** if  $A_1 \supset A_2 \supset \dots$  and then we define  $\lim_{n \rightarrow \infty} A_n = \bigcap_{i=1}^{\infty} A_i$ . In either case, we will write  $A_n \rightarrow A$ .

**1.4 Example.** Let  $\Omega = \mathbb{R}$  and let  $A_i = [0, 1/i)$  for  $i = 1, 2, \dots$ . Then  $\bigcup_{i=1}^{\infty} A_i = [0, 1)$  and  $\bigcap_{i=1}^{\infty} A_i = \{0\}$ . If instead we define  $A_i = (0, 1/i)$  then  $\bigcup_{i=1}^{\infty} A_i = (0, 1)$  and  $\bigcap_{i=1}^{\infty} A_i = \emptyset$ . ■

## 1.3 Probability

We will assign a real number  $\mathbb{P}(A)$  to every event  $A$ , called the **probability** of  $A$ .<sup>1</sup> We also call  $\mathbb{P}$  a **probability distribution** or a **probability measure**. To qualify as a probability,  $\mathbb{P}$  must satisfy three axioms:

**1.5 Definition.** A function  $\mathbb{P}$  that assigns a real number  $\mathbb{P}(A)$  to each event  $A$  is a **probability distribution** or a **probability measure** if it satisfies the following three axioms:

**Axiom 1:**  $\mathbb{P}(A) \geq 0$  for every  $A$

**Axiom 2:**  $\mathbb{P}(\Omega) = 1$

**Axiom 3:** If  $A_1, A_2, \dots$  are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

<sup>1</sup>It is not always possible to assign a probability to every event  $A$  if the sample space is large, such as the whole real line. Instead, we assign probabilities to a limited class of set called a  $\sigma$ -field. See the appendix for details.

There are many interpretations of  $\mathbf{P}(A)$ . The two common interpretations are frequencies and degrees of beliefs. In the frequency interpretation,  $\mathbf{P}(A)$  is the long run proportion of times that  $A$  is true in repetitions. For example, if we say that the probability of heads is  $1/2$ , we mean that if we flip the coin many times then the proportion of times we get heads tends to  $1/2$  as the number of tosses increases. An infinitely long, unpractical sequence of tosses whose limiting proportion tends to a constant is an idealization, much like the idea of a straight line in geometry. The degree-of-belief interpretation is that  $\mathbf{P}(A)$  measures an observer's strength of belief that  $A$  is true. In either interpretation, we require that Axioms 1 to 3 hold. The difference in interpretation will not matter much until we deal with statistical inference. There, the differing interpretations lead to two schools of inference: the frequentist and the Bayesian schools. We defer discussion until Chapter 13.

One can derive many properties of  $\mathbf{P}$  from the axioms, such as

$$\begin{aligned} \mathbf{P}(\emptyset) &= 0 \\ A \subset B &\implies \mathbf{P}(A) \leq \mathbf{P}(B) \\ 0 &< \mathbf{P}(A) < 1 \\ \mathbf{P}(A^c) &= 1 - \mathbf{P}(A) \\ A \cap B = \emptyset &\implies \mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B). \end{aligned} \quad (1.1)$$

A less obvious property is given in the following Lemma.

**1.6 Lemma.** For any events  $A$  and  $B$ ,

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB).$$

**PROOF.** Write  $A \cup B = (AB^c) \cup (AB) \cup (A^cB)$  and note that these events are disjoint. Hence, making repeated use of the fact that  $\mathbf{P}$  is additive for disjoint events, we see that

$$\begin{aligned} \mathbf{P}(A \cup B) &= \mathbf{P}\left((AB^c) \cup (AB) \cup (A^cB)\right) \\ &= \mathbf{P}(AB^c) + \mathbf{P}(AB) + \mathbf{P}(A^cB) \\ &= \mathbf{P}(AB^c) + \mathbf{P}(AB) + \mathbf{P}(A^cB) - \mathbf{P}(AB) + \mathbf{P}(AB) \\ &= \mathbf{P}\left((AB^c) \cup (AB)\right) + \mathbf{P}\left((A^cB) \cup (AB)\right) - \mathbf{P}(AB) \\ &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB). \quad \blacksquare \end{aligned}$$

**1.7 Example.** Two coin tosses. Let  $H_1$  be the event that heads occurs on toss 1 and let  $H_2$  be the event that heads occurs on toss 2. If all outcomes are

equally likely, then  $\mathbf{P}(H_1 \cup H_2) = \mathbf{P}(H_1) + \mathbf{P}(H_2) - \mathbf{P}(H_1 \cap H_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$ .

■

**1.8 Theorem (Continuity of Probabilities).** *If  $A_n \rightarrow A$  then*

$$\mathbf{P}(A_n) \rightarrow \mathbf{P}(A)$$

as  $n \rightarrow \infty$ .

**PROOF.** Suppose that  $A_n$  is monotone increasing so that  $A_1 \subset A_2 \subset \dots$ . Let  $A = \bigcap_{n=1}^{\infty} A_n = \bigcup_{i=1}^{\infty} B_i$ . Define  $B_1 = A_1$ ,  $B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}$ ,  $B_3 = \{\omega \in \Omega : \omega \in A_3, \omega \notin A_2, \omega \notin A_1\}$ , ... It can be shown that  $B_1, B_2, \dots$  are disjoint,  $A_n = \bigcup_{i=1}^n A_i = \bigcup_{i=1}^n B_i$  for each  $n$  and  $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$ . (See exercise 1.) From Axiom 3,

$$\mathbf{P}(A_n) = \mathbf{P}\left(\bigcup_{i=1}^n B_i\right) = \sum_{i=1}^n \mathbf{P}(B_i)$$

and hence, using Axiom 3 again,

$$\lim_{n \rightarrow \infty} \mathbf{P}(A_n) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbf{P}(B_i) = \sum_{i=1}^{\infty} \mathbf{P}(B_i) = \mathbf{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbf{P}(A). \quad \blacksquare$$

## 1.4 Probability on Finite Sample Spaces

Suppose that the sample space  $\Omega = \{\omega_1, \dots, \omega_n\}$  is finite. For example, if we toss a die twice, then  $\Omega$  has 36 elements:  $\Omega = \{(i, j) : i, j \in \{1, \dots, 6\}\}$ . If each outcome is equally likely, then  $\mathbf{P}(A) = |A|/36$  when  $|A|$  denotes the number of elements in  $A$ . The probability that the sum of the dice is 11 is  $2/36$  since there are two outcomes that correspond to this event.

If  $\Omega$  is finite and if each outcome is equally likely, then

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|},$$

which is called the **uniform probability distribution**. To compute probabilities, we need to count the number of points in an event  $A$ . Methods for counting points are called **combinatorial methods**. We won't delve into these in any great detail. We will, however, need a few facts from counting theory that will be useful later. Given  $n$  objects, the number of ways of ordering



- **[The Smoking Book for free](#)**
- [Genius Squad \(Genius, Book 2\) pdf, azw \(kindle\)](#)
- [Encyclopedia of Cold War Espionage, Spies, and Secret Operations pdf](#)
- [Veil of Night online](#)
  
- <http://www.1973vision.com/?library/The-Smoking-Book.pdf>
- <http://cavalldecartro.highlandagency.es/library/Hope-for-Film--From-the-Frontline-of-the-Independent-Cinema-Revolutions.pdf>
- <http://dadhoc.com/lib/US-Patrol-Torpedo-Boats--New-Vanguard--Volume-148-.pdf>
- <http://anvilpr.com/library/The-Devil-Never-Sleeps--and-Other-Essays.pdf>