

Contents

Preface

vii

1	Probability Theory (i): One Random Variable	1
1.1	Introduction	1
1.2	Discrete Random Variables, Definitions	3
1.2.1	Probability Distributions and Parameters	3
1.2.2	Independence	6
1.3	Six Important Discrete Probability Distributions	8
1.3.1	One Bernoulli Trial	8
1.3.2	The Binomial Distribution	8
1.3.3	The Hypergeometric Distribution	10
1.3.4	The Uniform Distribution	13
1.3.5	The Geometric Distribution	14
1.3.6	The Negative Binomial and the Generalized Geometric Distributions	16
1.3.7	The Poisson Distribution	17
1.3.8	Approximations	18
1.4	The Mean of a Discrete Random Variable	19
1.5	The Variance of a Discrete Random Variable	21
1.6	General Moments of a Probability Distribution	23
1.7	The Probability-Generating Function	24
1.8	Continuous Random Variables	26
1.9	The Mean, Variance, and Median of a Continuous Random Variable	28

1.9.1	Definitions	28
1.9.2	Chebyshev's Inequality	29
1.10	Five Important Continuous Distributions	30
1.10.1	The Uniform Distribution	30
1.10.2	The Normal Distribution	31
1.10.3	The Normal Approximation to a Discrete Distribution	32
1.10.4	The Exponential Distribution	34
1.10.5	The Gamma Distribution	36
1.10.6	The Beta Distribution	37
1.11	The Moment-Generating Function	38
1.12	Events	41
1.12.1	What Are Events?	41
1.12.2	Complements, Unions, and Intersections	42
1.12.3	Probabilities of Events	43
1.12.4	Conditional Probabilities	44
1.12.5	Independence of Events	46
1.13	The Memoryless Property of the Geometric and the Exponential Distributions	48
1.14	Entropy and Related Concepts	49
1.14.1	Entropy	49
1.14.2	Relative Entropy	50
1.14.3	Scores and Support	50
1.15	Transformations	51
1.16	Empirical Methods	53
2	Probability Theory (ii): Many Random Variables	62
2.1	Introduction	62
2.2	The Independent Case	64
2.3	Generating Functions	66
2.3.1	Properties of Probability-Generating Functions	66
2.3.2	Properties of Moment-Generating Functions	68
2.4	The Dependent Case	70
2.4.1	Covariance and Correlation	70
2.4.2	The Multinomial Distribution	71
2.4.3	The Multivariate Normal Distribution	72
2.5	Marginal Distributions	73
2.6	Conditional Distributions	75
2.7	Expected Values of Functions of Many Random Variables	80
2.8	Asymptotic Distributions	83
2.9	Indicator Random Variables	83
2.9.1	Definitions	83
2.9.2	Example: Sequencing EST Libraries	84
2.10	Derived Random Variables (i): Sums, Averages, and Minima	87

2.10.1	Sums and Averages	87
2.10.2	The Minimum of n Random Variables	90
2.11	Derived Random Variables (ii): The Maximum of n Random Variables	92
2.11.1	Distributional Properties: Continuous Random Variables	92
2.11.2	Distributional Properties: Discrete Random Variables	94
2.11.3	An Asymptotic Formula for the Distribution of $X_{\max}$	97
2.12	Order Statistics	99
2.12.1	Definition	99
2.12.2	Example: The Uniform Distribution	101
2.12.3	The Sample Median	102
2.13	Transformations	103
3	Statistics (i): An Introduction to Statistical Inference	111
3.1	Introduction	111
3.2	Classical and Bayesian Methods	112
3.3	Classical Estimation Methods	113
3.3.1	Unbiased Estimation	114
3.3.2	Confidence Intervals	114
3.3.3	Biased Estimators	116
3.4	Classical Hypothesis Testing	118
3.4.1	General Principles	118
3.4.2	P -Values	122
3.4.3	Power Calculations	123
3.5	Hypothesis Testing: Examples	124
3.5.1	Example 1. Testing for a Mean: the One-Sample Case	124
3.5.2	Example 2. The Two-Sample t -Test	125
3.5.3	Example 3. Tests on Variances	129
3.5.4	Example 4. Testing for the Parameters in a Multinomial Distribution	129
3.5.5	Example 5. Association tests.	130
3.6	Likelihood Ratios, Information, and Support	133
3.7	Hypothesis Testing Using a Maximum as Test Statistic	134
3.7.1	The Normal Distribution	134
3.7.2	P -values for the Maximum of Geometric Random Variables	136
3.8	Nonparametric Alternatives to the One-Sample and Two-Sample t -Tests	138
3.8.1	The Two-Sample Permutation Test	139
3.8.2	The Mann-Whitney Test	142
3.8.3	The Wilcoxon Signed-Rank Test	143

3.8.4	What is Assumed in a Non-Parametric Test? . . .	145
3.9	The Bayesian Approach to Hypothesis Testing	146
3.10	The Bayesian Approach to Estimation	148
3.11	Multiple Testing	149
3.12	Combining the Results of Several Experiments	151
4	Stochastic Processes (i): Poisson Processes and Markov Chains	155
4.1	The Homogeneous Poisson Process and the Poisson Distribution	155
4.2	The Poisson and the Binomial Distributions	158
4.3	The Poisson and the Gamma Distributions	159
4.4	The Pure-Birth Process	159
4.5	Introduction to Finite Markov Chains	161
4.6	Transition Probabilities and the Transition Probability Matrix	162
4.7	Markov Chains with Absorbing States	164
4.8	Markov Chains with No Absorbing States	164
4.8.1	Stationary Distributions	165
4.8.2	Example	166
4.9	The Graphical Representation of a Markov Chain	167
4.10	Modeling	168
5	The Analysis of One DNA Sequence	174
5.1	Shotgun Sequencing	174
5.1.1	Introduction	174
5.1.2	Contigs	175
5.1.3	Anchored Contigs	179
5.2	Modeling DNA	183
5.3	Modeling Signals in DNA	185
5.3.1	Introduction	185
5.3.2	Weight Matrices: Independence	185
5.3.3	Markov Dependencies	186
5.3.4	Maximal Dependence Decomposition	187
5.4	Long Repeats	188
5.5	r -Scans	192
5.6	The Analysis of Patterns	196
5.6.1	Introduction	196
5.6.2	Counting conventions	197
5.6.3	Notation and assumptions	198
5.7	Overlaps Counted	198
5.7.1	Number of Occurrences	198
5.7.2	Approximations to the Distribution of $Y_1(N)$	201
5.7.3	Distance Between Occurrences	203
5.7.4	Beginning at the Origin	205

5.8	Overlaps Not Counted	207
5.8.1	General Comments	207
5.8.2	Distance Between Recurrences	208
5.8.3	Number of Recurrences	209
5.9	Motifs	211
6	The Analysis of Multiple DNA or Protein Sequences	220
6.1	Two Sequences: Frequency Comparisons	220
6.2	Alignments	222
6.3	Simple Tests for Significant Similarity in an Alignment	223
6.4	Alignment Algorithms for Two Sequences	228
6.4.1	Introduction	228
6.4.2	Gapped Global Comparisons and Dynamic Programming Algorithms	230
6.4.3	Fitting One Sequence into Another Using a Linear Gap Model	233
6.4.4	Local Alignments with a Linear Gap Model	234
6.4.5	Other Gap Models	235
6.4.6	Limitations of the Dynamic Programming Alignment Algorithms	237
6.5	Protein Sequences and Substitution Matrices	238
6.5.1	Introduction	238
6.5.2	BLOSUM Substitution Matrices	239
6.5.3	PAM Substitution Matrices	244
6.5.4	A Simple Symmetric Evolutionary Matrix	248
6.6	Multiple Sequences	250
7	Stochastic Processes (ii): Random Walks	257
7.1	Introduction	257
7.2	The Simple Random Walk	259
7.2.1	Introduction	259
7.3	The Difference Equation Approach	260
7.3.1	Absorption Probabilities	260
7.3.2	Mean Number of Steps Taken Until the Walk Stops	261
7.4	The Moment-Generating Function Approach	262
7.4.1	Introduction and Wald's identity	262
7.4.2	Absorption Probabilities	263
7.4.3	Mean Number of Steps Until the Walk Stops	264
7.4.4	An Asymptotic Case	265
7.5	General Walks	266
7.6	General Walks: Asymptotic Theory	268
7.6.1	Introduction	268
7.6.2	The Renewal Theorem	268
7.6.3	Unrestricted Walks	269
7.6.4	Restricted Walks	271

8	Statistics (ii): Classical Estimation Theory	275
8.1	Introduction	275
8.2	Criteria for "Good" Estimators	275
8.3	Maximum Likelihood Estimation	277
8.3.1	The Discrete Case	277
8.3.2	The Continuous Case	279
8.3.3	Invariance Property	281
8.3.4	Asymptotic Properties	281
8.3.5	Many Parameters	285
8.4	Other Methods of Estimation	285
8.4.1	Introduction	285
8.4.2	The Method of Moments	285
8.4.3	Least Squares and Multiple Regression	287
8.5	Multivariate Methods	291
8.5.1	Introduction	291
8.5.2	Parameter Estimation	291
8.5.3	Principal Components	292
8.6	Bootstrap Methods: Estimation and Confidence Intervals	295
8.6.1	Introduction	295
8.6.2	The "Plug-in" Concept	295
8.6.3	Classical Estimation Methods	296
8.6.4	Bootstrap Estimation Methods	297
8.6.5	Bootstrap Confidence Intervals	299
8.6.6	Examples	299
9	Statistics (iii): Classical Hypothesis Testing Theory	304
9.1	Introduction	304
9.2	Simple Fixed-Sample-Size Tests	304
9.2.1	The Likelihood Ratio	304
9.2.2	The Neyman-Pearson Lemma	305
9.2.3	Example. Sequence matching.	307
9.3	Classical Hypothesis Testing: Composite Fixed-Sample-Size Tests	308
9.3.1	Introduction	308
9.3.2	Parameter Spaces and the Likelihood Ratio λ	308
9.3.3	Example: t -tests	309
9.4	The $-2 \log \lambda$ Approximation	313
9.4.1	Theory	313
9.4.2	Examples	315
9.5	The Analysis of Variance (ANOVA)	318
9.5.1	Introduction	318
9.5.2	From t to F : Sums of Squares and the F Statistic	319
9.5.3	One-way Fixed Effects ANOVA	320
9.5.4	The Two-Way Fixed Effects ANOVA	323
9.5.5	Multi-Way Fixed Effects ANOVAs	326

9.5.6	The 2^m Design, Confounding, and Fractional Replication	326
9.5.7	Random and Mixed Effects Models	328
9.6	Multivariate Methods	329
9.6.1	Introduction	329
9.6.2	One-sample T^2 Tests	330
9.6.3	Two-sample T^2 Tests	331
9.6.4	Optimal Linear Functions: Discriminant Functions	332
9.7	ANOVA: the Repeated Measures Case	332
9.8	Bootstrap Methods: the Two-sample t -test	334
9.9	Sequential Analysis	336
9.9.1	The Sequential Probability Ratio Test	336
9.9.2	The Power Function for a Sequential Test	338
9.9.3	The Mean Sample Size	340
9.9.4	The Effect of Boundary Overshoot	342
10	BLAST	345
10.1	Introduction	345
10.2	The Comparison of Two Aligned Sequences	346
10.2.1	Introduction	346
10.2.2	The BLAST Random Walk	346
10.2.3	Parameter Calculations	347
10.2.4	The Choice of a Score	349
10.2.5	Bounds and Approximations for the BLAST P -Value	351
10.2.6	The Normalized and the Bit Scores	354
10.2.7	The Number of High-Scoring Excursions	355
10.2.8	The Karlin-Altschul Sum Statistic	356
10.3	The Comparison of Two Unaligned Sequences	358
10.3.1	Introduction	358
10.3.2	Theoretical and Empirical Background	359
10.3.3	Edge Effects	361
10.3.4	Multiple Testing	363
10.4	The Comparison of a Query Sequence Against a Database	364
10.5	Printouts	365
10.5.1	Example	366
10.5.2	A More Complicated Example	369
10.6	Minimum Significance Lengths	372
10.6.1	A Correct Choice of n	372
10.6.2	An Incorrect Choice of n	374
10.7	BLAST: A Parametric or a Non-parametric Test?	375
10.8	Gapped BLAST and PSI BLAST	376
10.8.1	Gapped BLAST	376
10.8.2	PSI BLAST	378
10.9	Relation to Sequential Analysis	380

11 Stochastic Processes (iii): Markov Chains	384
11.1 Introduction	384
11.2 Markov Chains with No Absorbing States	385
11.2.1 Introduction	385
11.2.2 Convergence to the Stationary Distribution	385
11.2.3 Stationary Distributions: A Numerical Example	386
11.2.4 Reversibility and Detailed Balance	386
11.3 Higher-Order Markov Dependence	389
11.3.1 Testing for Higher-Order Markov Dependence	389
11.3.2 Testing for a Uniform Stationary Distribution	389
11.4 Patterns in Sequences with First-Order Markov Dependence	390
11.5 Markov Chain Monte Carlo	392
11.5.1 The Hastings-Metropolis Algorithm	392
11.5.2 Gibbs Sampling	393
11.5.3 Simulated Annealing	395
11.6 Markov Chains with Absorbing States	398
11.6.1 Theory	398
11.6.2 Examples	399
11.7 Continuous-Time Markov Chains	403
11.7.1 Definitions	403
11.7.2 Time-Dependent Solutions	404
11.7.3 The Stationary Distribution	404
11.7.4 Detailed Balance	405
11.7.5 Exponential Holding Times	405
11.7.6 The Embedded Chain	406
12 Hidden Markov Models	409
12.1 What is a Hidden Markov Model?	409
12.2 Three Algorithms	411
12.2.1 The Forward and Backward Algorithms	411
12.2.2 The Viterbi Algorithm	413
12.2.3 The Estimation Algorithms	414
12.3 Applications	417
12.3.1 Modeling Protein Families	417
12.3.2 Multiple Sequence Alignments	419
12.3.3 Pfam	421
12.3.4 Gene Finding	422
13 Gene Expression, Microarrays, and Multiple Testing	430
13.1 Introduction	430
13.1.1 Introduction to Microarrays	430
13.1.2 Microarray Data	432
13.1.3 Sources of Bias and Variation	438
13.2 The Statistical Analysis of Microarray Data: One Gene	443

13.2.1	Introduction	443
13.2.2	Determining Whether a Gene is Expressed	443
13.2.3	Testing for Differential Expression	444
13.3	Differential Expression – Multiple Genes	449
13.3.1	Introduction	449
13.3.2	Ranked lists	449
13.3.3	The Choice of Statistic	450
13.3.4	Confidence Measures	451
13.3.5	The Family-Wise Error Rate (FWER)	454
13.3.6	The False Discovery Rate (FDR)	459
13.3.7	The ANOVA Approach: Many Genes	467
13.3.8	Comparing Two Groups by Discriminant Analysis	470
13.4	Principal Components and Microarrays	471
13.5	Clustering Methods	471
13.5.1	Hierarchical Clustering	471
13.5.2	Other Forms of Clustering	472
14	Evolutionary Models	475
14.1	Models of Nucleotide Substitution	475
14.2	Discrete-Time Models	476
14.2.1	The Jukes–Cantor Model	476
14.2.2	The Kimura Models	478
14.2.3	Further Generalizations of the Kimura Models	479
14.2.4	The Felsenstein Models	480
14.2.5	The HKY Model	482
14.2.6	Other Models	483
14.2.7	The Reversibility Criterion	483
14.2.8	The Simple Symmetric Amino Acid Model	484
14.3	Continuous-Time Models	484
14.3.1	The Continuous-Time Jukes–Cantor Model	485
14.3.2	The Continuous-Time Kimura Model	489
14.3.3	The Continuous-Time Felsenstein Model	492
14.3.4	The Continuous-Time HKY Model	493
14.3.5	Continuous-Time Amino Acid Model	493
14.3.6	A Remark About Bias	494
15	Phylogenetic Tree Estimation	497
15.1	Introduction	497
15.2	Distances	499
15.3	Tree Reconstruction: The Ultrametric Case	501
15.4	Tree Reconstruction: the Neighbor-Joining Approach	505
15.5	Inferred Distances	509
15.6	Tree Reconstruction: Parsimony	511
15.7	Tree Estimation: Maximum Likelihood	512
15.8	Example	517

15.9	Modeling, Estimation, and Hypothesis Testing	522
15.9.1	Estimation and Hypothesis Testing	522
15.9.2	Bootstrapping and Phylogenies	523
15.9.3	Assumptions and Problems	528
15.9.4	Phylogenetic Models and Hypothesis Testing	531

Appendix A Basic Notions in Biology **537**

Appendix B Mathematical Formulae and Results **540**

B.1	Numbers and Intervals	540
B.2	Sets and Set Notation	541
B.3	Factorials	541
B.4	Binomial Coefficients	542
B.5	The Binomial Theorem	542
B.6	Permutations and Combinations	542
B.7	Limits	543
B.8	Asymptotics	544
B.9	Stirling's Approximation	546
B.10	Entropy as Information	546
B.11	Infinite Series	547
B.12	Taylor Series	549
B.13	Uniqueness of Taylor Series	553
B.14	Laurent Series	553
B.15	Numerical Solutions of Equations	554
B.16	Statistical Differentials	554
B.17	The Gamma Function	555
B.18	Proof by Induction	556
B.19	Linear Algebra and Matrices	557

Appendix C Computational Aspects of the Binomial and Generalized Geometric Distribution Functions **558**

Appendix D BLAST: Sums of Normalized Scores **560**

References	561
Author Index	581
Subject Index	586