

STATISTICS: METHODS AND APPLICATIONS

A COMPREHENSIVE REFERENCE FOR SCIENCE, INDUSTRY, AND DATA MINING

TABLE OF CONTENTS

Table of Contents	iii
Detailed Table of Contents	v
Preface	xv

1. Elementary Concepts in Statistics.....	1
2. Basic Statistics and Tables	15

STATISTICAL METHODS (in alphabetical order)

3. ANOVA/MANOVA.....	41
4. Association Rules	57
5. Boosting Trees.....	65
6. Canonical Analysis.....	69
7. CHAID Analysis.....	77
8. Classification and Regression Trees (CART)	81
9. Classification Trees.....	95
10. Cluster Analysis.....	115
11. Correspondence Analysis	127
12. Data Mining Techniques	141
13. Discriminant Function Analysis.....	155
14. Distribution Fitting.....	165
15. Experimental Design (Industrial DOE)	179

16. Factor Analysis and Principal Components	231
17. General Discriminant Analysis (GDA)	241
18. General Linear Models (GLM).....	245
19. General Regression Models (GRM).....	277
20. Generalized Additive Models (GAM).....	291
21. Generalized Linear/Nonlinear Models (GLZ).....	297
22. Log Linear Analysis of Frequency Tables.....	311
23. Machine Learning	317
24. Multivariate Adaptive Regression Splines (MARSplines).....	327
25. Multidimensional Scaling (MDS)	335
26. Multiple Linear Regression.....	341
27. Neural Networks.....	349
28. Nonlinear Estimation	365
29. Nonparametric Statistics.....	381
30. Partial Least Squares (PLS)	393
31. Power Analysis	405
32. Process Analysis	419
33. Quality Control Charts.....	443
34. Reliability/Item Analysis.....	457
35. Structural Equation Modeling	465
36. Survival/Failure Time Analysis.....	471
37. Text Mining	481
38. Time Series/Forecasting	489
39. Variance Components and Mixed Model ANOVA/ANCOVA	521
 Graphical Techniques	 531
Statistical Glossary.....	555
Distribution Tables	759
References Cited.....	777
Index.....	813

STATISTICS: METHODS AND APPLICATIONS

A COMPREHENSIVE REFERENCE FOR SCIENCE, INDUSTRY, AND DATA MINING

DETAILED TABLE OF CONTENTS

Table of Contents	iii
Detailed Table of Contents	v
Preface	xv
1. Elementary Concepts in Statistics	1
What are Variables?	3
What is Statistical Significance (p -Value)?	5
How to Measure the Magnitude (Strength) of Relations between Variables	8
Common "General Format" of Most Statistical Tests	9
How the Level of Statistical Significance is Calculated	9
Why the Normal Distribution is Important	10
2. Basic Statistics and Tables	15
Descriptive Statistics	17
Correlations	18
t -Test for Independent Samples	25
t -Test for Dependent Samples	26
Breakdown: Descriptive Statistics by Groups	28
Frequency Tables	31
Crosstabulation and Stub-and-Banner Tables	32

STATISTICAL METHODS (in alphabetical order)

3. ANOVA/MANOVA	41
Overview	43
Complex Designs	46
Analysis of Covariance (ANCOVA)	48
Multivariate Designs: MANOVA/MANCOVA	49
Contrast Analysis and Post-hoc Tests	50
Assumptions and Effects of Violating Assumptions	51
Methods for Analysis of Variance	54

4. Association Rules	57
Overview	59
Computational Procedures and Terminology	60
Tabular Representation of Associations	61
Graphical Representation of Associations	62
Interpreting and Comparing Results	63
5. Boosting Trees	65
Gradient Boosting Trees	67
The Problem of Overfitting; Stochastic Gradient Boosting	67
Stochastic Gradient Boosting Trees and Classification	68
6. Canonical Analysis	69
General Ideas	71
Sum Scores	71
Canonical Roots/Variates	71
Number of Roots	72
Extraction of Roots	72
Computational Methods and Results	72
Assumptions	75
7. CHAID Analysis	77
Basic Tree-Building Algorithm: CHAID and Exhaustive CHAID	79
General Computation Issues of CHAID	79
CHAID, CART, and QUEST	80
8. Classification and Regression Trees (CART)	81
Overview	83
Computational Details	87
Computational Formulas	92
9. Classification Trees	95
Overview	97
Characteristics of Classification Trees	98
Computational Methods	105
A Brief Comparison of Classification Tree Programs	113
10. Cluster Analysis	115
Statistical Significance Testing	117
Area of Application	117
Joining (Tree Clustering)	117
Two-Way Joining	120
k -Means Clustering	121
EM (Expectation Maximization) Clustering	122
Finding the Right Number of Clusters in k -Means and EM Clustering: V-Fold Cross-Validation	123

11. Correspondence Analysis	127
Overview	129
Supplementary Points	134
Multiple Correspondence Analysis (MCA)	136
Burt Tables	138
12. Data Mining Techniques	141
Overview	143
Crucial Concepts in Data Mining	144
Data Warehousing	149
On-Line Analytic Processing (OLAP)	150
Exploratory Data Analysis (EDA) and Data Mining Techniques	150
13. Discriminant Function Analysis	155
Computational Approach	157
Stepwise Discriminant Analysis	158
Interpreting a Two-Group Discriminant Function	158
Discriminant Functions for Multiple Groups	159
Assumptions	161
Classification	162
14. Distribution Fitting	165
Fit of the Observed Distribution	167
Types of Distributions	167
15. Experimental Design (Industrial DOE)	179
Overview	181
$2^{(k-p)}$ Fractional Factorial Designs at 2 Levels	185
$2^{(k-p)}$ Maximally Unconfounded and Minimum Aberration Designs	194
$3^{(k-p)}$, Box-Behnken, and Mixed 2- and 3-Level Factorial Designs	197
Central Composite and Non-Factorial Response Surface Designs	201
Latin Square Designs	206
Taguchi Methods: Robust Design Experiments	207
Mixture Designs and Triangular Surfaces	211
Designs for Constrained Surfaces and Mixtures	218
Constructing D- and A-Optimal Designs for Surfaces and Mixtures	221
Special Topics	226
16. Factor Analysis and Principal Components	231
Factor Analysis as a Data Reduction Method	233
Factor Analysis as a Classification Method	236
Other Issues and Statistics	239

17. General Discriminant Analysis (GDA)	241
Specifying Models for Predictor Variables and Predictor Effects	243
Stepwise and Best-Subset Analyses	243
Desirability Profiling of Posterior Classification Probabilities	243
A Note of Caution for Models with Categorical Predictors, and Other Advanced Techniques	244
The Use of Categorical Predictor Variables	244
18. General Linear Models (GLM)	245
Overview: The General Linear Model	247
Computations for Solving the Multiple Regression Equation	248
Extension of Multiple Regression to the General Linear Model	250
Types of Analyses	253
Estimation and Hypothesis Testing	265
Six Types of Sums of Squares	266
19. General Regression Models (GRM)	277
Overview: The Need for Simple Models	279
Model Building in General Stepwise Regression	279
Types of Analyses	280
Building the Whole Model	286
Building Models via Stepwise Regression	287
Building Models via Best-Subset Regression	289
20. Generalized Additive Models (GAM)	291
Overview	293
Generalized Additive Models	294
A Word of Caution	296
21. Generalized Linear/Nonlinear Models (GLZ)	297
Overview	299
Computational Approach	300
Types of Analyses	302
Model Building	309
Interpretation of Results and Diagnostics	310
22. Log Linear Analysis of Frequency Tables	311
Two-Way Frequency Tables	313
Multi-Way Frequency Tables	314
Log Linear Model	314
Goodness of Fit	315
Automatic Model Fitting	315

23. Machine Learning	317
Support Vector Machines Overview.....	319
Naive Bayes Classifier Overview	321
<i>k</i> -Nearest Neighbors Overview.....	323
24. Multivariate Adaptive Regression Splines (MARSplines)	327
Overview	329
The MARSplines Algorithm.....	332
The MARSplines Model.....	332
25. Multidimensional Scaling (MDS)	335
Logic of Multidimensional Scaling	337
Computational Approach.....	337
How Many Dimensions to Specify?	338
Interpreting the Dimensions.....	339
Applications.....	340
MDS and Factor Analysis.....	340
26. Multiple Linear Regression.....	341
Overview	343
Computational Approach.....	343
Assumptions, Limitations, Practical Considerations	346
27. Neural Networks	349
Artificial Neural Networks	351
Biological Inspiration	351
Why Use Neural Networks?	351
Nonparametric (Black Box) Models.....	352
Single Neuron Model.....	353
Multi-Neuron Model.....	354
Neural Network Complexity.....	355
Universal Approximator	355
Multilayer Perceptrons (MLP).....	355
Radial Basis Function (RBF).....	356
Labeled and Unlabeled Data.....	357
Neural Networks for Regression.....	358
Neural Networks for Classification	358
Other Tasks Performed by Neural Networks.....	358
Learning by Example – Training.....	359
Neural Network Training.....	359
The Problem of Overfitting	360
Network Complexity Revisited	362
Validating Neural Networks and Generalization.....	362
Recommended Textbooks	363

28. Nonlinear Estimation	365
Overview	367
Estimating Linear and Nonlinear Models	367
Common Nonlinear Regression Models	368
Nonlinear Estimation Procedures.....	373
Evaluating the Fit of the Model	378
29. Nonparametric Statistics.....	381
Overview	383
Brief Overview of Nonparametric Methods	384
When to Use Which Method.....	385
Nonparametric Correlations.....	386
2x2 Tables, Chi/V/Phi Square, McNemar, Fisher Exact.....	386
Correlations (Spearman, Kendall Tau, Gamma).....	387
Wald-Wolfowitz Runs Test	388
Kolmogorov-Smirnov Two-Sample Test.....	388
Mann-Whitney U Test	389
Kruskal-Wallis ANOVA by Ranks and Median Test	389
Sign Test	389
Wilcoxon Matched Pairs Test	390
Friedman ANOVA and Kendall Concordance	390
Cochran Q Test	390
Ordinal Descriptive Statistics.....	391
30. Partial Least Squares (PLS).....	393
Overview	395
Computational Approach	396
Training (Analysis) and Verification (Cross-Validation) Samples.....	397
Types of Analyses.....	397
Distance Graphs	404
31. Power Analysis	405
Overview	407
Power Analysis and Sample Size Calculation in Experimental Design.....	407
Noncentrality Interval Estimation and the Evaluation of Statistical Models	415
32. Process Analysis	419
Sampling Plans.....	421
Process (Machine) Capability Analysis	423
Gage Repeatability and Reproducibility	427
Non-Normal Distributions	430
Weibull and Reliability/Failure Time Analysis	433

33. Quality Control Charts	443
Overview	445
Establishing Control Limits	445
Common Types of Charts	446
Short Run Charts	447
Unequal Sample Sizes	448
Control Charts for Variables vs. Charts for Attributes	449
Control Charts for Individual Observations	449
Out-of-Control Process: Runs Tests	449
Operating Characteristic (OC) Curves	451
Process Capability Indices	451
Other Specialized Control Charts	453
34. Reliability/Item Analysis	457
Overview	459
Basic Ideas	459
Classical Testing Model	460
Reliability	460
Sum Scales	460
Cronbach's Alpha	461
Split-Half Reliability	461
Correction for Attenuation	462
Designing a Reliable Scale	462
35. Structural Equation Modeling	465
Overview	467
Basic Idea Behind Structural Modeling	467
Structural Equation Modeling and the Path Diagram	468
36. Survival/Failure Time Analysis	471
Overview	473
Censored Observations	473
Analytic Techniques	473
Life Table Analysis	473
Distribution Fitting	475
Kaplan-Meier Product-Limit Estimator	475
Comparing Samples	476
Regression Models	477
37. Text Mining	481
Overview	483
Some Typical Applications for Text Mining	483
Approaches to Text Mining	484

Issues and Considerations for “Numericizing” Text.....	485
Transforming Word Frequencies	486
Latent Semantic Indexing via Singular Value Decomposition	487
Incorporating Text Mining Results in Data Mining Projects	488
38. Time Series/Forecasting	489
Overview	491
Two Main Goals.....	491
Identifying Patterns in Time Series Data	491
ARIMA	494
Interrupted Time Series ARIMA	499
Exponential Smoothing.....	499
Classical Seasonal Decomposition (Census Method 1)	504
X-11 Census Method II Seasonal Adjustment	506
Distributed Lags Analysis Overview	510
Single Spectrum (Fourier) Analysis.....	511
Cross-Spectrum Analysis.....	512
Spectrum Analysis – Basic Notation and Principles.....	514
Fast Fourier Transformations (FFT)	518
39. Variance Components and Mixed Model ANOVA/ANCOVA	521
Overview.....	523
Estimation of Variance Components (Technical Overview)	524
Graphical Techniques	531
Categorized Graphs.....	533
Brushing	544
Smoothing Bivariate Distributions.....	545
Layered Compression	546
Projections of 3D Data Sets	546
Icon Plots	547
Data Reduction.....	552
Data Rotation (in 3D Space)	553
Statistical Glossary	555
A.....	557
B.....	565
C.....	582
D.....	602
E.....	610
F.....	617
G.....	620

H	626
I	630
J	636
K	638
L	642
M	655
N	671
O	677
P	682
Q	697
R	703
S	715
T	740
U	747
V	749
W	751
X	755
Y	756
Z	757
Distribution Tables	759
Standard Normal (Z) Table	761
Student's t Table	763
Chi-Square Table	765
F Distribution Tables	768
References Cited	777
Index	813