

Contents

Preface	vii
1 Empirical Research	1
1.1 AI Programs as Objects of Empirical Studies	2
1.2 Three Basic Research Questions	3
1.3 Answering the Basic Research Questions	4
1.4 Kinds of Empirical Studies	5
1.5 Data Analysis for Empirical Studies	5
1.6 A Prospective View of Empirical Artificial Intelligence	6
2 Exploratory Data Analysis	9
2.1 Data	9
2.1.1 Scales of Data	12
2.1.2 Transforming Data	12
2.1.3 Measurement Theory	13
2.2 Sketching a Preliminary Causal Model	14
2.3 Looking at One Variable	16
2.3.1 Visualizing One Variable	16
2.3.2 Statistics for One Variable	18
2.4 Joint Distributions	21
2.4.1 Joint Distributions of Categorical and Ordinal Variables	21
2.4.2 Contingency Tables for More than Two Variables	25
2.4.3 Statistics for Joint Distributions of Categorical Variables	27
2.4.4 Visualizing Joint Distributions of Two Continuous Variables	30
2.4.5 Statistics for Joint Distributions of Two Continuous Variables	36
2.4.6 The Sensitivity of Pearson's Correlation Coefficient to Outliers	39
2.5 Time Series	41
2.5.1 Visualizing Time Series	41

2.5.2	Smoothing	42
2.5.3	Statistics for Time Series	45
2.6	Execution Traces	47
2.6.1	Visualizing Execution Traces	47
2.6.2	Statistics for Execution Traces	49
3	Basic Issues in Experiment Design	51
3.1	The Concept of Control	52
3.1.1	What is an Extraneous Variable?	54
3.1.2	Control Conditions in MYCIN: A Case Study	56
3.2	Four Spurious Effects	59
3.2.1	Ceiling and Floor Effects	59
3.2.2	How to Detect Ceiling and Floor Effects	60
3.2.3	Bounding Performance	62
3.2.4	Regression Effects	62
3.2.5	Order Effects	63
3.3	Sampling Bias	66
3.4	The Dependent Variable	69
3.5	Pilot Experiments	70
3.6	Guidelines for Experiment Design	71
3.7	Tips for Designing Factorial Experiments	72
3.8	The Purposes of Experiments	74
3.9	Ecological Validity: Making Experiments Relevant	75
3.10	Conclusion	76
4	Hypothesis Testing and Estimation	79
4.1	Statistical Inference	79
4.2	Introduction to Hypothesis Testing	80
4.3	Sampling Distributions and the Hypothesis Testing Strategy	82
4.3.1	Sampling Distributions	83
4.3.2	How to Get Sampling Distributions	84
4.4	Tests of Hypotheses about Means	88
4.4.1	The Anatomy of the Z Test	88
4.4.2	Critical Values	91
4.4.3	p Values	92
4.4.4	When the Population Standard Deviation Is Unknown	92
4.4.5	When All Population Parameters Are Unknown	93

4.4.6	When N is Small: The t Test	93
4.4.7	Two-Sample t Test	94
4.4.8	The Paired Sample t Test	96
4.5	Hypotheses about Correlations	96
4.6	Parameter Estimation and Confidence Intervals	98
4.6.1	Confidence Intervals for μ When σ is Known	99
4.6.2	Confidence Intervals for μ When σ is Unknown	99
4.6.3	An Application of Confidence Intervals: Error Bars	100
4.6.4	How Big Should Samples Be?	102
4.6.5	Errors	104
4.7	Conclusion	108
4.8	Further Reading	108
5	Computer-Intensive Statistical Methods	109
5.1	Monte Carlo Tests	111
5.2	Bootstrap Methods	113
5.2.1	Bootstrap Sampling Distributions for Censored Data	114
5.2.2	Bootstrap Two-sample Tests	118
5.2.3	Bootstrap Confidence Intervals	119
5.3	Randomization Tests	122
5.3.1	A Randomization Version of the Two-Sample t Test.	124
5.3.2	A Randomization Version of the Paired Sample t Test.	124
5.3.3	A Randomization Test of Independence.	126
5.3.4	Randomization for a Robust Statistic: The Resistant Line.	127
5.4	Comparing Bootstrap and Randomization Procedures	129
5.5	Comparing Computer-intensive and Parametric Procedures	130
5.6	How Many Pseudosamples?	133
5.7	Jackknife and Cross Validation	133
5.8	An Illustrative Nonparametric Test: The Sign Test	133
5.9	Conclusion	135
5.10	Further Reading	135
6	Performance Assessment	137
6.1	Strategies for Performance Assessment	138
6.2	Comparisons to External Standards: The View Retriever	138
6.2.1	Introduction to Pairwise Comparisons of Means	140
6.2.2	Introduction to Analysis of Variance	142

6.2.3	An Analysis of Acker and Porter’s Data	143
6.2.4	Unplanned Pairwise Comparisons: Scheffé Tests	144
6.2.5	Unplanned Pairwise Comparisons: LSD Tests	145
6.2.6	Which Test? Interpretations of “Conservative”	146
6.3	Comparisons among Many Systems: The MUC-3 Competition	147
6.4	Comparing the Variability of Performance: Humans vs. the View Retriever	152
6.5	Assessing Whether a Factor Has Predictive Power	153
6.6	Assessing Sensitivity: MYCIN’s Sensitivity to Certainty Factor Accuracy.	154
6.7	Other Measures of Performance in Batches of Trials	155
6.8	Assessing Performance During Development: Training Effects in OTB	156
6.9	Cross-validation: An Efficient Training and Testing Procedure	159
6.10	Learning Curves	162
6.11	Assessing Effects of Knowledge Engineering with Retesting	163
6.12	Assessing Effects with Classified Retesting: Failure Recovery in Phoenix	164
6.12.1	Expected Frequencies from Other Sources	168
6.12.2	Heterogeneity, Independence, and Goodness-of-Fit Tests	169
6.12.3	Diminishing Returns and Overfitting in Retesting	171
6.13	Conclusion	172
6.14	Appendix to Chapter 6: Analysis of Variance and Contrast Analysis	174
6.14.1	One-Way Analysis of Variance	174
6.14.2	Contrasts, or Comparisons Revisited	178
7	Explaining Performance: Interactions and Dependencies	185
7.1	Strategies for Explaining Performance	186
7.2	Interactions among Variables: Analysis of Variance	186
7.2.1	Introduction to Two-Way Analysis of Variance	187
7.2.2	Two-way Analysis of Phoenix Data	188
7.2.3	Three-way Analysis of Phoenix Data	189
7.3	Explaining Performance with Analysis of Variance	193
7.3.1	Looking for No Effect	193
7.3.2	Getting a Clearer Picture by Reducing Variance	196
7.3.3	Explaining Nonlinear Effects: Transforming Data	197
7.3.4	Summary: Analysis of Variance	198
7.4	Dependencies among Categorical Variables: Analysis of Frequencies	199
7.5	Explaining Dependencies in Execution Traces	199
7.6	Explaining More Complex Dependencies	201

7.7	General Patterns in Three-way Contingency Tables	207
7.7.1	Complete Independence	209
7.7.2	One-factor Independence	210
7.7.3	Conditional Independence	211
7.7.4	Homogenous Association	212
7.8	Conclusion	212
7.9	Further Reading	213
7.10	Appendix to Chapter 7: Experiment Designs and Analyses	213
7.10.1	Two-Way Fixed Factorial Design Without Repeated Measures	214
7.10.2	A Numerical Example	219
7.10.3	Two-way Mixed Design Without Repeated Measures	220
7.10.4	One-Way Design with Repeated Measures	223
7.10.5	When Systems Learn	226
8	Modeling	229
8.1	Programs as Models: Executable Specifications and Essential Miniatures	231
8.2	Cost as a Function of Learning: Linear Regression	234
8.2.1	Introduction to Linear Regression	235
8.2.2	Lack of Fit and Plotting Residuals	237
8.3	Transforming Data for Linear Models	238
8.4	Confidence Intervals for Linear Regression Models	240
8.4.1	Parametric Confidence Intervals	240
8.4.2	Bootstrap Confidence Intervals	241
8.5	The Significance of a Predictor	243
8.6	Linear Models with Several Predictors: Multiple Regression	243
8.7	Standardized Regression Coefficients	244
8.8	A Model of Plan Adaptation Effort	246
8.9	Causal Models	249
8.9.1	Why Causal Modeling is Difficult	250
8.9.2	Regression Coefficients as Causal Strengths	252
8.10	Structural Equation Models	252
8.11	Conclusion	256
8.12	Further Reading	256
8.13	Appendix to Chapter 8: Multiple Regression	256
8.13.1	Normal Equations	258
8.13.2	Standardized Coefficients	258

8.13.3	Normal Equations for Standardized Variates	259
8.13.4	A Causal Interpretation of Regression: Path Diagrams	260
8.13.5	Regression Coefficients Are Partial	261
8.13.6	Testing the Significance of Predictors	263
9	Tactics for Generalization	265
9.1	Empirical Generalization	267
9.2	Theories and “Theory”	270
9.3	Tactics for Suggesting and Testing General Theories	272
9.4	A Theory about Task and Architecture Features	273
9.5	Bounding the Scope of Theories and Predicted Behavior	274
9.6	Noticing Analogous Features in the Literature	276
9.7	Which Features?	276
9.8	Finding the “Same” Behavior in Several Systems	277
9.9	The Virtues of Theories of Ill-Defined Behavior	278