Contents

Preface *ix*

Acknowledgments *xv*

---

1 Empirical Research *1*

1.1 AI Programs as Objects of Empirical Studies *2*

1.2 Three Basic Research Questions *4*

1.3 Answering the Basic Research Questions *5*

1.4 Kinds of Empirical Studies *6*

1.5 A Prospective View of Empirical Artificial Intelligence *9*

2 Exploratory Data Analysis *11*

2.1 Data *12*

2.2 Sketching a Preliminary Causal Model *18*

2.3 Looking at One Variable *20*

2.4 Joint Distributions *27*

2.5 Time Series *53*

2.6 Execution Traces *62*

3 Basic Issues in Experiment Design *67*

3.1 The Concept of Control *68*

3.2 Four Spurious Effects *79*

3.3 Sampling Bias *89*

3.4 The Dependent Variable *92*
3.5 Pilot Experiments *94*
3.6 Guidelines for Experiment Design *96*
3.7 Tips for Designing Factorial Experiments *97*
3.8 The Purposes of Experiments *100*
3.9 Ecological Validity: Making Experiments Relevant *101*
3.10 Conclusion *103*

4 Hypothesis Testing and Estimation *105*

4.1 Statistical Inference *106*
4.2 Introduction to Hypothesis Testing *106*
4.3 Sampling Distributions and the Hypothesis Testing Strategy *110*
4.4 Tests of Hypotheses about Means *117*
4.5 Hypotheses about Correlations *130*
4.6 Parameter Estimation and Confidence Intervals *132*
4.7 How Big Should Samples Be? *137*
4.8 Errors *140*
4.9 Power Curves and How to Get Them *143*
4.10 Conclusion *145*
4.11 Further Reading *146*

5 Computer-Intensive Statistical Methods *147*

5.1 Monte Carlo Tests *150*
5.2 Bootstrap Methods *153*
5.3 Randomization Tests *165*
5.4 Comparing Bootstrap and Randomization Procedures *175*
5.5 Comparing Computer-Intensive and Parametric Procedures *177*
5.6 Jackknife and Cross Validation *180*
5.7 An Illustrative Nonparametric Test: The Sign Test *180*
5.8 Conclusion *182*
5.9 Further Reading *183*

6 Performance Assessment *185*

6.1 Strategies for Performance Assessment *186*

6.2 Assessing Performance in Batches of Trials *187*

6.3 Comparisons to External Standards: The View Retriever *187*

6.4 Comparisons among Many Systems: The MUC-3 Competition *199*

6.5 Comparing the Variability of Performance: Humans vs. the View Retriever *205*

6.6 Assessing Whether a Factor Has Predictive Power *207*

6.7 Assessing Sensitivity: MYCIN's Sensitivity to Certainty Factor Accuracy *208*

6.8 Other Measures of Performance in Batches of Trials *210*

6.9 Assessing Performance During Development: Training Effects in OTB *211*

6.10 Cross-Validation: An Efficient Training and Testing Procedure *216*

6.11 Learning Curves *219*

6.12 Assessing Effects of Knowledge Engineering with Retesting *221*

6.13 Assessing Effects with Classified Retesting: Failure Recovery in Phoenix *223*

6.14 Diminishing Returns and Overfitting in Retesting *232*

6.15 Conclusion *233*

Appendix: Analysis of Variance and Contrast Analysis *235*

7 Explaining Performance: Interactions and Dependencies *249*

7.1 Strategies for Explaining Performance *250*

7.2 Interactions among Variables: Analysis of Variance *251*

7.3 Explaining Performance with Analysis of Variance *260*

7.4 Dependencies among Categorical Variables: Analysis of Frequencies *267*

7.5 Explaining Dependencies in Execution Traces *268*

7.6 Explaining More Complex Dependencies *270*

7.7 General Patterns in Three-Way Contingency Tables *279*

7.8 Conclusion *287*

7.9 Further Reading *287*

Appendix: Experiment Designs and Analysis *287*

8 Modeling *309*

8.1 Programs as Models: Executable Specifications and Essential Miniatures *312*

8.2 Cost as a Function of Learning: Linear Regression *316*

8.3 Transforming Data for Linear Models *321*

8.4 Confidence Intervals for Linear Regression Models *324*

8.5 The Significance of a Predictor *327*

8.6 Linear Models with Several Predictors: Multiple Regression *328*

8.7 A Model of Plan Adaptation Effort *332*

8.8 Causal Models *337*

8.9 Structural Equation Models *342*

8.10 Conclusion *347*

8.11 Further Reading *347*

Appendix: Multiple Regression *348*

9 Tactics for Generalization *359*

9.1 Empirical Generalization *362*

9.2 Theories and "Theory" *366*

9.3 Tactics for Suggesting and Testing General Theories *369*

9.4 Which Features? *375*

9.5 Finding the "Same" Behavior in Several Systems *376*

9.6 The Virtues of Theories of Ill-Defined Behavior *378*

References *385*

Index *395*