

Table of Contents
Empirical Methods for Artificial Intelligence
Paul R. Cohen

Preface

1 Empirical Research

- 1.1 AI Programs as Objects of Empirical Studies
- 1.2 Three Basic Research Questions
- 1.3 Answering the Basic Research Questions
- 1.4 Kinds of Empirical Studies
- 1.5 Data Analysis for Empirical Studies
- 1.6 A Prospective View of Empirical Artificial Intelligence

2 Exploratory Data Analysis

- 2.1 Data
 - 2.1.1 Scales of Data
 - 2.1.2 Transforming Data
 - 2.1.3 Measurement Theory
- 2.2 Sketching a Preliminary Causal Model
- 2.3 Looking at One Variable
 - 2.3.1 Visualizing One Variable
 - 2.3.2 Statistics for One Variable
- 2.4 Joint Distributions
 - 2.4.1 Joint Distributions of Categorical and Ordinal Variables
 - 2.4.2 Contingency Tables for More than Two Variables
 - 2.4.3 Statistics for Joint Distributions of Categorical Variables
 - An Easy and Useful Special Case: Two-by-two Table
 - 2.4.4 Visualizing Joint Distributions of Two Continuous Variables
 - Evidence of Independence in Scatterplots
 - Point Coloring to Find Potential Causal Factors
 - Fitting Functions to Data in Scatterplots
 - 2.4.5 Statistics for Joint Distributions of Two Continuous Variables
 - 2.4.6 The Sensitivity of Pearson's Correlation Coefficient to Outliers
 - Other Statistics for Joint Distributions of Variables
- 2.5 Time Series
 - 2.5.1 Visualizing Time Series
 - 2.5.2 Smoothing
 - 2.5.3 Statistics for Time Series
- 2.6 Execution Traces
 - 2.6.1 Visualizing Execution Traces
 - 2.6.2 Statistics for Execution Traces

3 Basic Issues in Experiment Design

- 3.1 The Concept of Control
 - 3.1.1 What is an Extraneous Variable?
 - 3.1.2 Control Conditions in MYCIN: A Case Study
- 3.2 Four Spurious Effects
 - 3.2.1 Ceiling and Floor Effects
 - 3.2.2 How to Detect Ceiling and Floor Effects
 - 3.2.3 Bounding Performance
 - 3.2.4 Regression Effects
 - 3.2.5 Order Effects
- 3.3 Sampling Bias
- 3.4 The Dependent Variable
- 3.5 Pilot Experiments
- 3.6 Guidelines for Experiment Design
- 3.7 Tips for Designing Factorial Experiments
- 3.8 The Purposes of Experiments
- 3.9 Ecological Validity: Making Experiments Relevant
- 3.10 Conclusion

4 Hypothesis Testing and Estimation

- 4.1 Statistical Inference
- 4.2 Introduction to Hypothesis Testing
- 4.3 Sampling Distributions and the Hypothesis Testing Strategy
 - 4.3.1 Sampling Distributions
 - 4.3.2 How to Get Sampling Distributions
 - Exact Sampling Distributions: The Sampling Distribution of the Proportion
 - Estimated Sampling Distributions: The Sampling Distribution of the Mean
 - The Standard Error of the Mean and Sample Size
- 4.4 Tests of Hypotheses about Means
 - 4.4.1 The Anatomy of the Z Test
 - 4.4.2 Critical Values
 - 4.4.3 p Values
 - 4.4.4 When the Population Standard Deviation Is Unknown
 - 4.4.5 When All Population Parameters Are Unknown
 - 4.4.6 When N is Small: The t Test
 - 4.4.7 Two-Sample t Test
 - 4.4.8 The Paired Sample t Test
- 4.5 Hypotheses about Correlations
- 4.6 Parameter Estimation and Confidence Intervals
 - 4.6.1 Confidence Intervals for μ When σ is Known
 - 4.6.2 Confidence Intervals for μ When σ is Unknown
 - 4.6.3 An Application of Confidence Intervals: Error Bars
 - 4.6.4 How Big Should Samples Be?

4.6.5 Errors

Power Curves and How to Get Them

4.7 Conclusion

4.8 Further Reading

5 Computer-Intensive Statistical Methods

5.1 Monte Carlo Tests

5.2 Bootstrap Methods

5.2.1 Bootstrap Sampling Distributions for Censored Data

Bootstrap Sampling Distributions for H_0 : Shift Method

Bootstrap Sampling Distributions for H_0 : Normal Approximation Method

5.2.2 Bootstrap Two-sample Tests

5.2.3 Bootstrap Confidence Intervals

5.3 Randomization Tests

5.3.1 A Randomization Version of the Two-Sample t Test

5.3.2 A Randomization Version of the Paired Sample t Test

Other Two Sample and Paired Sample Randomization Tests

5.3.3 A Randomization Test of Independence

5.3.4 Randomization for a Robust Statistic: The Resistant Line

5.4 Comparing Bootstrap and Randomization Procedures

5.5 Comparing Computer-intensive and Parametric Procedures

5.6 How Many Pseudosamples?

5.7 Jackknife and Cross Validation

5.8 An Illustrative Nonparametric Test: The Sign Test

5.9 Conclusion

5.10 Further Reading

6 Performance Assessment

6.1 Strategies for Performance Assessment

6.2 Comparisons to External Standards: The View Retriever

6.2.1 Introduction to Pairwise Comparisons of Means

6.2.2 Introduction to Analysis of Variance

6.2.3 An Analysis of Acker and Porter's Data

6.2.4 Unplanned Pairwise Comparisons: Scheffé Tests

6.2.5 Unplanned Pairwise Comparisons: LSD Tests

6.2.6 Which Test? Interpretations of "Conservative"

6.3 Comparisons among Many Systems: The MUC-3 Competition

6.4 Comparing the Variability of Performance: Humans vs. the View Retriever

6.5 Assessing Whether a Factor Has Predictive Power

6.6 Assessing Sensitivity: MYCIN's Sensitivity to Certainty Factor Accuracy

6.7 Other Measures of Performance in Batches of Trials

6.8 Assessing Performance During Development: Training Effects in OTB

6.9 Cross-validation: An Efficient Training and Testing Procedure

- 6.10 Learning Curves
- 6.11 Assessing Effects of Knowledge Engineering with Retesting
- 6.12 Assessing Effects with Classified Retesting: Failure Recovery in Phoenix
 - 6.12.1 Expected Frequencies from Other Sources
 - 6.12.2 Heterogeneity, Independence, and Goodness-of-Fit Tests
 - 6.12.3 Diminishing Returns and Overfitting in Retesting
- 6.14 Appendix to Chapter 6: Analysis of Variance and Contrast Analysis
 - 6.14.1 One-Way Analysis of Variance
 - Foundation of One-way Analysis of Variance
 - Hypothesis Testing with One-way Analysis of Variance
 - A Numerical Example
 - 6.14.2 Contrasts, or Comparisons Revisited
 - Orthogonal Contrasts
 - Testing the Significance of Orthogonal Contrasts
- 7 Explaining Performance: Interactions and Dependencies
 - 7.1 Strategies for Explaining Performance
 - 7.2 Interactions among Variables: Analysis of Variance
 - 7.2.1 Introduction to Two-Way Analysis of Variance
 - 7.2.2 Two-way Analysis of Phoenix Data
 - 7.2.3 Three-way Analysis of Phoenix Data
 - 7.3 Explaining Performance with Analysis of Variance
 - 7.3.1 Looking for No Effect
 - Negative and Positive Evidence for No Effect
 - Random Factors and Generalization
 - 7.3.2 Getting a Clearer Picture by Reducing Variance
 - 7.3.3 Explaining Nonlinear Effects: Transforming Data
 - 7.3.4 Summary: Analysis of Variance
 - 7.4 Dependencies among Categorical Variables: Analysis of Frequencies
 - 7.5 Explaining Dependencies in Execution Traces
 - 7.6 Explaining More Complex Dependencies
 - 7.7 General Patterns in Three-way Contingency Tables
 - 7.7.1 Complete Independence
 - 7.7.2 One-factor Independence
 - 7.7.3 Conditional Independence
 - 7.7.4 Homogenous Association
 - 7.8 Conclusion
 - 7.9 Further Reading
 - 7.10 Appendix to Chapter 7: Experiment Designs and Analyses
 - 7.10.1 Two-Way Fixed Factorial Design Without Repeated Measures
 - 7.10.2 A Numerical Example
 - 7.10.3 Two-way Mixed Design Without Repeated Measures
 - 7.10.4 One-Way Design with Repeated Measures

7.10.5 When Systems Learn

8 Modeling

8.1 Programs as Models: Executable Specifications and Essential Miniatures

8.2 Cost as a Function of Learning: Linear Regression

8.2.1 Introduction to Linear Regression

Parameters of the Regression Line

Variance Accounted for by the Regression Line

8.2.2 Lack of Fit and Plotting Residuals

8.3 Transforming Data for Linear Models

8.4 Confidence Intervals for Linear Regression Models

8.4.1 Parametric Confidence Intervals

8.4.2 Bootstrap Confidence Intervals

8.5 The Significance of a Predictor

8.6 Linear Models with Several Predictors: Multiple Regression

8.7 Standardized Regression Coefficients

8.8 A Model of Plan Adaptation Effort

8.9 Causal Models

8.9.1 Why Causal Modeling is Difficult

8.9.2 Regression Coefficients as Causal Strengths

8.10 Structural Equation Models

8.11 Conclusion

8.12 Further Reading

8.13 Appendix to Chapter 8: Multiple Regression

8.13.1 Normal Equations

8.13.2 Standardized Coefficients

8.13.3 Normal Equations for Standardized Variates

8.13.4 A Causal Interpretation of Regression: Path Diagrams

8.13.5 Regression Coefficients Are Partial

8.13.6 Testing the Significance of Predictors

9 Tactics for Generalization

9.1 Empirical Generalization

9.2 Theories and “Theory”

9.3 Tactics for Suggesting and Testing General Theories

9.4 A Theory about Task and Architecture Features

9.5 Bounding the Scope of Theories and Predicted Behavior

9.6 Noticing Analogous Features in the Literature

9.7 Which Features?

9.8 Finding the “Same” Behavior in Several Systems

9.9 The Virtues of Theories of Ill-Defined Behavior