Exploring and Analyzing Data

Course Description

The goal of this course is to provide students with an introduction to many different types of quantitative research methods and statistical techniques for analyzing data. We begin with a focus on measurement, inferential statistics, and causal inference. Then we will explore a range of statistical techniques and methods using the open-source statistics language, R. We will use many different statistics and techniques for analyzing and viewing data, with a focus on *applying* this knowledge to real-world data problems. Topics in quantitative techniques include descriptive and inferential statistics, sampling, experimental design, parametric and nonparametric tests of difference, ordinary least squares regression, general linear models. (3 units)

Prerequisites

At least one college-level statistics course or equivalent.

Course Assignments and Final Exam

There are three short "lab assignments" in this course. These will be completed in small groups. The final exam will cover all of the major topics in the course (methods, data management and analysis techniques, statistical analysis and interpretation).

- Lab assignments (3 total): 60%
- Final exam: 30%
- Participation 10%

Course Software

We will use the R statistical computing environment in this course. R is a language *and* environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, etc.) and graphical techniques, and is highly extensible. R is free to download and use on Windows, Mac OS, and most UNIX platforms. http://www.r-project.org/

Textbooks and Reader

This course will use a primary textbook and some online readings. All students are expected to do the readings before watching lectures and before all discussion sections. In many cases the lecture materials and the weekly discussion sessions will not make sense if you do not have knowledge of the examples and issues from the readings.

Field, Andy, Jeremy Miles, and Zoe Field. *Discovering Statistics Using R*. SAGE Publications, 2012.

<u>Online Reader</u>: Selected chapters from other textbooks as well as empirical articles and examples.

Syllabus

Week1: Introduction to Fundamental Concepts of Measurement

Lecture: Introduction: Measurement and Inference

Reading: Field, Miles, and Field, chapter 1 – 1.6

- Variable types (ordinal, metric, nominal, etc.)
- Constructs vs. variables
- Operationalization of constructs
- Validity and reliability of constructs
- Precision and accuracy
- Knowledge claims from constructs

Week 2: Probability and Sampling: Where Did These Data Come from Anyway?

Lecture: Non-Probability and Probability Sampling

Readings:

Bernard, Russell H. *Social Research Methods: Qualitative and Quantitative Approaches*, 2nd ed. SAGE Publications, 2000, chapter 5 “Sampling.”

Freedman, David, Robert Pisani, Roger Purves. *Statistics*, 4th ed. W. W. Norton & Company, Inc., 2007, chapter 13 “What Are the Chances?” and chapter 14 “More about Chance.”

- Probability rules (addition, multiplication)
- Chance error vs. bias
- Quota sampling

- Purposive sampling
- Convenience sampling
- Samples vs. census of behavior data
- Sampling frames
- Simple random samples
- Systematic random samples
- Stratified random sampling

Week 3: Essentials of Statistics

Lecture: Simple Statistical Models and Using Statistical Models to Test Research Questions

Reading: Field, Miles, and Field, chapter 1.7, chapter 2-2.5

- Measures of central tendency
- The normal curve
- Sums of squares, variance, standard deviation
- Building statistical models
- Standard error
- Confidence intervals
- Central limit theorem, law of large numbers

Week 4: Working in the R Environment: Data Preparation, Variable Recoding, Data Manipulation and Introduction to Exploring Data with Graphs

Lecture: Working in R, Basics and Exploring with Graphs

Reading: Field, Miles, and Field, chapter 3

- Data preparation in R
- Recoding and error-checking
- Viewing data in R
- Recoding, creating and testing composite measures, aggregations over time/people/other units
- Frequency distributions
- Scatterplots
- Histograms

LAB ASSIGNMENT #1 DISTRIBUTED

Week 5: Exploring Data with Graphs and Plots in R

Lecture: Exploring Data with Graphs

Reading: Field, Miles, and Field, chapter 4

- Properties of good graphs
- Ggplot2
- More with scatterplots
- Boxplots
- Density plots
- Bar charts, error bars
- Line graphs
- . . . and why pie charts are the devil's tool

Week 6: The Logic of Hypothesis Testing

Lecture: Causality, Posing and Testing Hypotheses, Statistical Significance

Reading: Field, Miles, and Field, chapter 2.6

- Propositions and hypotheses
- Conditions of causality
- Causation and casual paths (direct, reciprocal, indirect)
- z-scores, z-test, and area under normal curve
- Null and alternative hypotheses
- One- and two-tailed hypothesis testing
- Critical values
- p-values
- Type I and type II errors
- Measuring practical significance

Week 7: Exploring Assumptions

Lecture: Exploring Assumptions

Reading: Field, Miles, and Field, chapter 5

- Assumptions of normality
- Testing if a distribution is normal
- Testing for homogeneity of variance
- Dealing with outliers
- Non-normal and unequal variances
- Transformations (non-normal distributions)

LAB ASSIGNMENT #2 DISTRIBUTED

Week 8: Tests of Association

Lecture: Measures of Association with Metric Data

Reading: Field, Miles, and Field, chapter 6-6.5.6, 6.6, 6.9 and chapter 18 – 18.6

- Covariance
- Correlation, Pearson’s r, SD lines
- Interpretation of scatterplots (homogeneity, nonlinear, censored/limited data)
- Measures of association with nominal data
  - Chi-square distribution and the chi-square test of independence
  - Goodman-Kruskal gamma

Week 9: Comparing Two Means: Independent and Dependent tests

Lecture: Looking at Differences

Reading: Field, Miles, and Field, chapter 9 and chapter 15-15.5

- z and t-distributions (large and small data samples)
- Degrees of freedom
- Independent sample t-tests
- Repeated sample t-test (dependent test)
- Practical significance testing (Cohen’s d)
- Nonparametric tests of difference
  - Wilcoxon signed-rank test
  - Mann-Whitney
  - Kruskal-Wallis

Week 10: ANOVA

Lecture: ANOVA

Reading: Field, Miles, and Field, chapter 10

- F-ratio
- f-distribution and f-test
- Assumptions of ANOVA
- Correcting for multiple comparisons (Bonferroni, Sidak, LSD, etc.)

Week 11: Experiments

Lecture: Basics of Experimentation

Readings:

Freedman, David, Robert Pisani, Roger Purves. *Statistics*, 4th ed. W. W. Norton & Company, Inc. 2007, chapter 1 “Controlled Experiments.”

Dahlia K. Remler, Gregg G. Van Ryzin. Research Methods in Practice: Strategies for Description and Causation, chapter 13 “Natural and Quasi Experiments.”

- Experimental design vs. observational studies
- Controls and treatments
- Variable independence
- Random assignment vs. random sample
- “True” experiments
- Pretest-posttest, post only, Solomon 4-group, etc.
- Internal validity, external validity, ecological validity
- Quasi-experimental designs
- Associational nonexperimental designs
- Online field experiments

LAB ASSIGNMENT #3 DISTRIBUTED

Week 12: OLS Regression

Lecture: OLS Regression and Diagnostics

Reading: Field, Miles, and Field, chapter 7-7.5

- Linear model, ordinary least squares solution
- Root mean square, scatter
- Correlation and regression, coefficient of determination (R2)
- Heteroskedasticity
- Multicollinearity (variance inflation factors)
- Transformations (non-normal data)

Week 13: Multiple and Nested Regression

Lecture: Multiple Regression and Nested Regression

Reading: Field, Miles, and Field, chapter 7.6-7.12

- Nested modeling; real-world data examples
- Teasing out explanatory factors
- F-change testing between models, R2 comparisons

Week 14: General Linear Model Approaches: Logistic Regression

Lecture: Maximum Likelihood versus OLS

Reading: Field, Miles, and Field, chapter 8 -8.8

- Maximum likelihood method of estimation
- Logistic regression
    - Binary response variables
    - Interpreting logistic regression
    - Coefficients, odds ratios, model statistics

FINAL EXAM DISTRIBUTED

Week 15: Course Review, Interviews with Data Science Professionals, Final Exam

Lecture: Interviews, Course Review Material (asynch only this week)