Term Information

Effective Term	Autumn 2014

General Information

Course Bulletin Listing/Subject Area	Statistics
Fiscal Unit/Academic Org	Statistics - D0694
College/Academic Group	Arts and Sciences
Level/Career	Undergraduate
Course Number/Catalog	3202
Course Title	Introduction to Statistical Inference for Data Analytics
Transcript Abbreviation	Intr Stat Inf DA
Course Description	The course covers foundational inferential methods for learning about populations from samples, including point and interval estimation, and the formulation and testing of hypotheses. Statistical theory is introduced to justify the approaches. The course emphasizes challenges that arise when applying classical ideas to big data, partially through the use of computational and simulation techniques.
Semester Credit Hours/Units	Fixed: 4

Offering Information

Length Of Course	14 Week
Flexibly Scheduled Course	Never
Does any section of this course have a distance education component?	No
Grading Basis	Letter Grade
Repeatable	No
Course Components	Lecture, Recitation
Grade Roster Component	Lecture
Credit Available by Exam	No
Admission Condition Course	No
Off Campus	Never
Campus of Offering	Columbus

Prerequisites and Exclusions

Prerequisites/Corequisites	Stat 3201
Exclusions	

Cross-Listings

Cross-Listings	

Subject/CIP Code

Subject/CIP Code	27.0501
Subsidy Level	Baccalaureate Course
Intended Rank	Sophomore, Junior

Requirement/Elective Designation

Required for this unit's degrees, majors, and/or minors

Course Details

Course goals or learning objectives/outcomes

- Describe the role of a parameter in a statistical model and its relationship to observed data
- Use data to estimate and describe uncertainty about the parameters of a statistical model
- Translate scientific hypotheses about a population into mathematical statements about parameters in a statistical model
- Formulate statistical procedures to test a hypothesis about parameters in a statistical model, and interpret the results in both statistical and application-specific terms
- Explain the difference between statistical and practical significance in massive data settings
- Appreciate the effect of missing data on statistical inference
- Evaluate and compare different statistical procedures for answering the same question

Content Topic List

- Statistical models and parameters
- point and interval estimation
- effects of missing data
- formulating statistical hypotheses
- tests for means, variances and proportions
- interpreting and explaining the results of statistical tests
- properties of hypothesis tests

Attachments

- 3202_Syllabus.pdf

(Syllabus. Owner: Hans,Christopher M)

Comments

- This is a required course for the proposed major in Data Analytics. *(by Craigmile,Peter F on 10/11/2013 03:19 PM)*

Workflow Information

Status	User(s)	Date/Time	Step
Submitted	Hans,Christopher M	10/09/2013 02:50 PM	Submitted for Approval
Approved	Craigmile,Peter F	10/13/2013 06:09 PM	Unit Approval
Approved	Hadad,Christopher Martin	10/14/2013 06:51 AM	College Approval
Pending Approval	Vankeerbergen,Bernadette Chantal Nolen,Dawn Jenkins,Mary Ellen Bigler Hogle,Danielle Nicole Hanlin,Deborah Kay	10/14/2013 06:51 AM	ASCCAO Approval

Statistics 3202

Introduction to Statistical Inference for Data Analytics

4-semester-hour course

Prerequisite: Stat 3201 (Introduction to Probability for Data Analytics)

Exclusions:

Class distribution: Three 55-minute lectures and one 55-minute recitation per week

Course Description and Learning Outcomes

The course covers foundational inferential methods for learning about populations from samples, including point and interval estimation, and the formulation and testing of hypotheses. Statistical theory is introduced to justify the approaches. The course emphasizes challenges that arise when applying classical ideas to big-data, partially through the use of computational and simulation techniques.

Upon successful completion of the course, students will be able to

1. Describe the role of a parameter in a statistical model and its relationship to observed data
2. Use data to estimate and describe uncertainty about the parameters of a statistical model
3. Translate scientific hypotheses about a population into mathematical statements about parameters in a statistical model
4. Formulate statistical procedures to test a hypothesis about parameters in a statistical model, and interpret the results in both statistical and application-specific terms
5. Explain the difference between statistical and practical significance in massive data settings
6. Appreciate the effect of missing data on statistical inference
7. Evaluate and compare different statistical procedures for answering the same question

Required Text and Other Course Materials

The required textbook for the course is *Mathematical Statistics with Applications* (7th edition) by Wackerly, Mendenhall and Sheaffer. The book is available for purchase at the official University bookstore (ohiostate.bkstore.com) and elsewhere online. The book is available on reserve in the 18th Avenue Library.

Students will be required to use the R[1] software environment for statistical computing and graphics. R can be downloaded for free at http://www.r-project.org. Instructions for using the software will be given in class. Many students prefer to use RStudio, an IDE designed for use with R. RStudio is available for free at http://www.rstudio.com.

Recitation

Weekly recitations will reinforce the material introduced in lecture by having students solve both analytical and computational problems. The recitations will focus on details of problem solving and computational implementation.

Assignments

Homework will be assigned (approximately) weekly, will be due on the dates announced in class and will be graded. Assignments will consist of a mix of several problems selected from the textbook, problems motivated by data analytics applications, and small computer simulation problems.

Two group projects will be assigned during the semester, as described briefly below:

Project 1: Small-group project on the topic of estimation. The instructor will formulate a question based on an interesting data set and will propose various estimators that might be used to help answer the question. Groups will be assigned to investigate and compare different aspects of the various estimators using simulation. Group presentations of the results will be given in class, and a written report of the results must be submitted. Each group will also provide written feedback for one other group.

Project 2: Small-group project on the topic of testing. The instructor will formulate a question based on an interesting data set and will specify various hypotheses and tests that might be used to answer the question. Groups will be assigned to investigate one particular aspect of the testing problem using simulation. Group presentations of the results will be given in class, and a written report of the results must be submitted. Each group will also provide written feedback for one other group.

---

[1] For information on the use of R in data analytics, see:

- http://www.revolutionanalytics.com/why-revolution-r/whitepapers/r-is-hot.php
- http://techcrunch.com/2012/10/27/big-data-right-now-five-trendy-open-source-technologies/
- http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html
- http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/

Exams

There will be two in-class midterms that cover material from lecture, the assigned readings and homework.

A final examination will be given during the university's examination period.

Grading Information

The final course grade will be based on homework assignments, two projects, two midterms and a comprehensive final examination. The weights for each component of the grade are:

Homework	Project 1	Project 2	Midterm 1	Midterm 2	Final Exam
20%	10%	10%	20%	20%	20%

Outline of topics

1. Introduction to inference
    a. General ideas: parameters, statistics, population, sample
    b. Classical inference (e.g., estimate a parameter) vs. modern inference from big data (e.g., infer relatedness on networks, infer evolution of network connections over time, estimate structures such as trees, etc.)
    c. Inference frameworks: frequentist, Bayesian
2. Sampling distributions
    a. Review of sampling distributions
    b. Asymptotic distributions of known form (e.g., CLT)
    c. Sampling distributions in complex settings (e.g., contingency tables, networks)
    d. *Simulation activity*: sampling distributions in complex settings
3. Estimation
    a. Topics in point estimation: bias, variance, sufficiency, completeness
    b. Interval estimation: estimation of variances, methods of constructing intervals
    c. Bootstrap estimation of variances, confidence intervals
    d. *Simulation activity:* Examining bootstrap approaches
        i. Parametric and nonparametric bootstrap
        ii. Comparison with existing approaches
    e. Missing data

        i. Effects on properties of estimators
        ii. Approaches to handling missing data
    f. *Simulation activity*: Exploring the effect of missing data
        i. Differences between patterns of missing data
        ii. Comparison of approaches to handling missing data
4. Testing
    a. Formulation of hypotheses
    b. Classical methods of forming tests (e.g., likelihood ratio tests)
    c. Concept of null distribution
    d. Methods of specifying null distribution: parametric form, nonparametric tests, permutation, simulation
    e. Type I and type II errors
    f. Statistical and practical significance
    g. Interpretation of p-values
    h. Multiple testing issues
    i. *Simulation activities*:
        i. Simulation-based evaluation of type I and type II error rates
        ii. Simulation of null distribution and comparison with other methods of specifying null
        iii. Examination of power under various alternatives
        iv. Comparison of multiple testing procedures
5. Introduction to Modeling
    a. What is a statistical model?
    b. Estimation in statistical models
    c. Model comparison
6. Case studies: estimation in complex settings
    a. Discussion of several examples of estimation for big data problems
    b. Examples involving estimation and/or testing for a complex data set – address the issues discussed throughout the semester

Statement on Academic Misconduct

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term “academic misconduct” includes all forms of student academic

misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

Special Accommodations

Students with disabilities that have been certified by the Office for Disability Services will be appropriately accommodated and should inform the instructor as soon as possible of their needs. The Office for Disability Services is located in 150 Pomerene Hall, 1760 Neil Avenue; telephone 292-3307, TDD 292-0901; http://www.ods.ohio-state.edu/.