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	3201
Course Title	Introduction to Probability for Data Analytics
Transcript Abbreviation	Intr Prob for DA
Course Description	An introduction to probability and its role in statistical methods for data analytics. Equal emphasis is placed on analytical and simulation-based methods for quantifying uncertainty. Approaches to assessing the accuracy of simulation methods are discussed. Students should have some prior knowledge of basic programming. Applications of probability and sampling to big-data settings are discussed.
Semester Credit Hours/Units	Fixed: 3

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
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	Math 1152 or 1161.xx or 1172 or 1181 or equiv, and CSE Placement Level A or equiv; or permission of the instructor.
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

- Quantify uncertainty about events using mathematical descriptions of probability
- Quantify uncertainty about events using simulation methods
- Assess the quality and accuracy of simulation-based descriptions of uncertainty
- Update a description of uncertainty based on new information
- Identify appropriate probability models for experiments/data and summarize expected outcomes from such models
- Use correlation and conditional expectation to describe the relationship between two random variables
- Quantify uncertainty about summary statistics for large data sets

Content Topic List

- Definitions and interpretations of probability
- conditional probability
- random variables
- simulating random events
- assessing accuracy of simulation methods
- correlation
- sampling distributions
- central limit theorem
- sampling methods

Attachments

- 3201_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 03:18 PM)*

Workflow Information

Status	User(s)	Date/Time	Step
Submitted	Hans,Christopher M	10/09/2013 02:49 PM	Submitted for Approval
Approved	Craigmile,Peter F	10/13/2013 06:08 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 3201

Introduction to Probability for Data Analytics

3-semester-hour course

Prerequisite: Math 1152 or 1161.xx or 1172 or 1181 or equiv, and CSE Placement Level A or equiv; or permission of the instructor.

Exclusions:

Class distribution: Three 55-minute lectures per week

Course Description and Learning Outcomes

An introduction to probability and its role in statistical methods for data analytics. Equal emphasis is placed on analytical and simulation-based methods for quantifying uncertainty. Approaches to assessing the accuracy of simulation methods are discussed. Students should have some prior knowledge of basic programming. Applications of probability and sampling to big-data settings are discussed.

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

1. Quantify uncertainty about events using mathematical descriptions of probability
2. Quantify uncertainty about events using simulation methods
3. Assess the quality and accuracy of simulation-based descriptions of uncertainty
4. Update a description of uncertainty based on new information
5. Identify appropriate probability models for experiments/data and summarize expected outcomes from such models
6. Use correlation and conditional expectation to describe the relationship between two random variables.
7. Quantify uncertainty about summary statistics for large data sets

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

[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/

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.

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 consisting of a simulation of simple random processes to estimate probabilities and describe uncertainty (example settings: card games, simple genetics examples, finite population sampling, etc.). An investigation of the quality and accuracy of the simulation must be reported. 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. Suggested topics will be provided; group-initiated topics are allowed with approval of the instructor.

Project 2: Small group project to investigate the sampling distribution of a statistic that is a nonlinear function of data (motivated by a big data application). Groups to report on both small and large sample properties. Suggested topics will be provided; group-initiated topics are allowed with approval of the instructor.

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. The role of uncertainty in data analysis and decision-making
2. Introduction to the R statistical computing environment
3. Numerical and graphical methods for characterizing data
4. Introduction to probability
    a. Counting methods
    b. Conditional Probability
5. Simulation-based methods for estimating probabilities
    a. Assessing the accuracy of simulation methods
6. Random variables and their distributions
    a. Useful continuous and discrete distributions
    b. Properties of random variables
    c. Methods for simulating random variables
7. Relationships between random variables
    a. Transformation techniques
    b. Joint distribution of two random variables
    c. Conditional distributions
    d. Bivariate normal distribution
8. Random samples and statistics
    a. Simulating the sampling distribution of a statistic
    b. Samples from normal populations, χ^2 and *t* distributions
    c. Central Limit Theorem
9. Simple random and stratified sampling

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/.