

ISTA 116: Statistical Foundations for the Information Age

Mihai Surdeanu

Last Revised January 14, 2014

1 Course Information

1.1 Course Description

Data is everywhere in the modern world. Each of us has access to data on virtually any topic imaginable via the internet, and public and private institutions are making use of data in decision-making more than ever before. This trend will only accelerate. However, access to data is not the same thing as access to information. The purpose of this course is to develop some of the foundational skills needed to consume data and create information. The main theme in the course is understanding the variability inherent in data, and the inherent uncertainty associated with conclusions drawn from data.

The first part of the course is dedicated to descriptive statistics: taking a data set, visualizing and summarizing its key features. How often does each value occur, what are the “typical” values, how much variation does it contain, and what are the relationships among the variables? In the second part, we develop the foundations of inferential statistics, namely the basic principles of probability and the behavior of random samples. Finally, we discuss how to evaluate how strong the evidence is for simple claims about the data, such as “homeopathic headache medicine reduces pain more effectively than sugar”, or “children’s income is correlated with parent’s income”.

Throughout the semester, the lab component of the class will introduce computational techniques (using the statistical language R) to make our machines do the

tedious work of counting, arithmetic, and generating random numbers. That is, until they rebel and enslave us.

1.2 Prerequisites

MATH 109C or MATH 112 or placement beyond the level of College Algebra.

1.3 Locations and Times

Lectures:

Monday/Wednesday 4:30pm - 5:45pm in Saguaro Hall, Room 202

Labs:

Tues 12:00 – 1:50 in ECE 229 (lab instructor: Kate)

Tues 2:00 – 3:50 in ECE 229 (lab instructor: Sarah)

Tues 4:00 – 5:50 in ECE 229 (lab instructor: Sarah)

1.4 Texts

The main text for the lectures will be:

Griffiths, Dawn. (2008). *Head First Statistics*. Sebastopol, CA: O'Reilly. <http://universityofarizona.worldcat.org.ezproxy1.library.arizona.edu/title/head-first-statistics/oclc/297556511> (available for free electronically through UA library)

There is no text for the lab component, but the lab instructors will provide pdf notes each week. In addition to the lab notes, the following is a good reference for R. **We highly recommend reading the first several chapters during the first few weeks:**

Owen, W. J. (2010). *The R Guide, ver. 2.5*. Department of Mathematics and Computer Science, University of Richmond, Richmond, VA. <http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf> (available for free electronically)

For those who would like to go a bit deeper (in either statistics or R, but especially R), the following book has good coverage:

Cohen, Y. and Cohen, J. (2008). *Data and Statistics with R*. Wiley & Sons. (available free through the UA library).

1.5 Instructor Information

Mihai Surdeanu

Email: msurdeanu@email.arizona.edu

Office: Gould-Simpson 811

Office Hours: Wed 1 – 2 and by appointment

1.6 Lab Instructors

TA: Yekaterina (Kate) Kharitonova

Email: ykk@email.arizona.edu

Office: Gould-Simpson 927

Office Hours: Tue 9:30 – 10:30 and by appointment

SL: Sarah Menchaca

Email: menchacasarah@email.arizona.edu

Office: Gould-Simpson 856

Office Hours: Thu 3 – 5

2 Course Policies and Grading

2.1 Grading

Grades are based on homework, quizzes, one midterm, and a comprehensive final exam. The grading scheme is as follows:

Component	Weight
Homework	300 pts
Web Assignments	200 pts
Term Paper	100 pts
Midterm Exam	200 pts
Final Exam	200 pts
Paper Presentation (optional)	25 pts
Total	1025 pts

Grade	Point Range
A	> 900
B	775-899
C	650-774
D	525-649
E	< 525

Grade Disputes

Disputes about grades on a particular assignment or exam will be entertained for two weeks from the day the assignment is returned, or 1 day before grades are due, whichever is sooner. These will be resolved by re-grading the entire assignment or exam. Note that this can result in a lower grade in the event that new mistakes are discovered. The final exam will be graded and made available for review by students within 24 hours of its completion, to allow time for any requested regrades.

The final exam will be graded and made available for review by students within 24 hours of its completion, to allow time for any requested regrades.

No negotiations about individual students' letter grades will be entertained once final grades are assigned, except as permitted by the policy stated above.

2.2 Homework Assignments

There will be 7 homework assignments, consisting of a mixture of paper-and-pencil and computer-based problems. Individual assignments are worth 50 points, and the

lowest score is dropped. Assignments are graded for correctness; however, genuine effort is nearly always worth partial credit.

Collaboration Policy

Students are encouraged to work together, both in class / lab / office hours and otherwise, to understand problems and general approaches for solutions. However, **final write-ups of solutions must be done individually**. Any collaboration that takes place outside section or office hours must be identified, along with the nature of the collaboration (e.g., “we worked together”, “she helped me”, “I helped him”). **Copying another person’s answers, work or code is not permitted, regardless of collaboration status, and will be treated as a case of academic dishonesty.**

Late Policy

Assignments are due electronically via D2L by the stated deadline. Permission for an extension must be granted by the lab instructor *in advance* of the deadline in order to receive full credit for a late assignment. The first request by a given student is likely to be granted; the probability decreases with each subsequent request. No homework will be accepted once solutions are posted online.

Calculation of Final Homework Grade

Each student’s final homework grade will be based on the six highest homework scores, **provided all assignments are completed**. “Completion” is judged by the lab instructors, and is based on an honest attempt to answer nearly all of the questions. If more than one assignment is incomplete, no assignment will be dropped.

2.3 Web Assignments

Web assignments will be due in D2L (generally) during weeks when there is no written homework due. These are intended to provide practice with theoretical statistical concepts, divorced from computational concerns. These assignments can be redone

up to 3 times; the highest score will count. Students are encouraged to make at least their first attempt without any reference materials.

2.4 Exams

One in-class midterm exam will be given during the lecture on March 12th (last lecture before Spring Recess). A comprehensive final exam will be given between May 12th and May 15th (exact date and time TBA). The format of the exams will be mixed, and may include multiple choice, short verbal answer, and “math problems”. Each exam is graded out of 200 pts. The higher of the two scores will be upweighted at the end of the semester to be out of 225 points, and the lower downweighted to be out of 175.

Permitted Materials

Students will be allowed to use one double-sided $8\frac{1}{2} \times 11$ ” page of notes for the midterm, and two pages for the final. A simple, self-contained hand calculator may also be used. **No other electronic devices of any kind, including cell phones, iPods, laptops, Borg implants, etc., will be permitted during the exams.**

Make-up Policy

Exams may only be made up in case of a serious, unanticipated emergency. The student or an appointed proxy must contact the instructor in advance of the exam. Clearly worded documentation, from a doctor, dean, etc., demonstrating physical inability to take the exam at the appointed time, will be required to grant a make-up appointment. **No make-ups will be granted for personal reasons, including travel or personal hardship.** Any make-up exams that are granted must be scheduled for after the original time.

2.5 Term Paper

There will be one short (roughly 4-6 page) paper due on May 4th. The purpose of the paper is to find a data set on your own and say something meaningful about it

using the concepts and techniques learned in the course.

Students who want to present their findings to the class will have the opportunity to do so during the last week of classes for extra credit; however, the number of presentation slots is likely to be smaller than the number of students in the class, so sign-ups will be first come, first served. The research topic and data set must be defined and approved before sign-ups may occur.

3 University Policies

Classroom Behavior

Students are expected to behave respectfully toward each other and to the instructor and TAs. Disrespectful behavior includes the use of cell phones or other electronic devices in the classroom during class hours.

The Arizona Board of Regents Student Code of Conduct is here: <http://dos.web.arizona.edu/uapolicies/scc5308abcd.html#sccphilosophy>

ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to oneself. See: <http://policy.web.arizona.edu/~policy/threaten.shtml>.

Special Needs and Accommodations

Students who need special accommodation or services should contact the

Disability Resources Center
1224 East Lowell Street, Tucson, AZ 85721
(520) 621-3268
FAX (520) 621-9423
email: uadrc@email.arizona.edu
web: <http://drc.arizona.edu/>.

You must register and request that the Center or DRC send official notification of your accommodations needs as soon as possible. Please plan to meet with the instructor by appointment or during office hours to discuss accommodations and how the course requirements and activities may impact your ability to fully participate. The need for accommodations must be documented by the appropriate office.

Student Code of Academic Integrity

Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work/exercises must be the product of independent effort unless otherwise instructed. Students are expected

to adhere to the UA Code of Academic Integrity as described in the UA General Catalog. See: <http://dos.web.arizona.edu/uapolicies/>.

Confidentiality of Student Records

See <http://www.registrar.arizona.edu/ferpa/default.htm>

Subject to Change Statement

Information contained in this syllabus, other than the grade and absence policy, may be subject to change by the instructor, with advance notice.

4 Schedule

4.1 Tentative Schedule of Lecture Topics

Topic	Approx. Dates	Readings
Introduction	Jan 15	Course Syllabus Ch. 1 thru “A Tale of Two Charts”
Categorical Data	Jan 22, 27, 29	Ch. 1 thru “Categories vs. Numbers” Lecture Notes
Frequency Distributions	Feb 3	finish Ch. 1
Central Tendency	Feb 5	Ch. 2
Variability	Feb 10, 12	Ch. 3
Inference and Prediction	Feb 17, 19	Ch. 11 until “Let’s Estimate...” Lecture Notes
Sampling	Feb 24	Ch. 10 thru “How to Choose...”
Probability Basics	Feb 26, Mar 3	Ch. 4 until “Conditions Apply”
Conditional Probability	Mar 5	finish Ch. 4
Catch-up Day	Mar 10	—
MIDTERM EXAM	Mar 12	—
Conditional Probability, ctd.	Mar 24, 26	finish Ch. 4
Discrete Random Variables	Mar 31	Ch. 5 until “Fat Dan Changed...”
Permutations and Combinations	Apr 2	Ch. 6
The Geometric and Binomial Distributions	Apr 7	Ch. 7: Ch. 7 from “Who Wants to Win...” until “Statsville Cinema...”
Hypothesis Testing: Binary Data	Apr 9, 14	Ch. 13 until “Let’s Conduct Another...”
Permutation Tests	Apr 16	Lecture Notes
Bootstrap Confidence Intervals	Apr 21, 23	Lecture Notes
Correlation & Regression	Apr 28, 30	Ch. 15 Lecture Notes
Student Presentations	May 5, 7	—
FINAL EXAM	Between May 12 – 15	—

4.2 Tentative Schedule of Lab Topics

Date	Topics
Jan 21	RStudio and R Fundamentals
Jan 28	Importing and Manipulating Data
Feb 4	Contingency Tables, Bar Plots, and Conditioning
Feb 11	Histograms & Density Curves
Feb 18	Summary Statistics, Visualization, Boxplots
Feb 25	Sampling and <code>replicate()</code>
Mar 4	Probability Spaces
Mar 11	<i>Midterm Review</i>
Mar 25	Expectation and Variance
Apr 1	Binomial Distribution
Apr 8	Hypothesis-testing Exercises
Apr 15	Monte Carlo Simulation
Apr 22	Permutation Tests
Apr 29	Bootstrap Resampling
May 6	Correlation/Regression Simulations

4.3 Homework and Web Assignments

All assignments are due in the D2L dropbox by 11:59 P.M. on the indicated day.

Assignment	Due Date
HW 0 + WA 1	Jan 19
WA 2	Jan 26
HW 1	Feb 2
HW 2	Feb 9
WA 3	Feb 16
HW 3	Feb 23
WA 4	Mar 2
HW 4	Mar 9
	<i>Midterm + Recess</i>
WA 5	Mar 30
HW 5	Apr 6
WA 6	Apr 13
HW 6	Apr 20
WA 7	Apr 27
HW 7	May 4th