

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

SI 544 Introductory Statistics and Data Analysis

(Preliminaries)

Lada Adamic

School of Information,
University of Michigan

Jan. 3rd, 2007

Outline

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

motivation

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

Dick DeVaux:

"We haven't evolved to be statisticians. Our students who think statistics is an unnatural subject are right. This isn't how humans think naturally. But it is how humans think rationally. And it is how scientists think. This is the way we must think if we are to make progress in understanding how the world works and, for that matter, how we ourselves work."

motivation

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

Gary King (Department of Government, Harvard University):

"Statisticians will rule the world."

(When discussing the opportunities that availability of massive data sets will present for addressing questions in social science.)

how I see things

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- There is lots of interesting data
- We need to describe what is going on
For this we need **descriptive statistics**
 - need to summarize
 - need to visualize
- We need to tell whether what we are seeing is an actual trend, or part of the random “noise”
To this end we need to
 - Understand probability
 - Understand probability distributions (what is the likelihood that this would occur by chance?)

example (from Dick DeVaux)

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- A town has two hospitals
 - Large hospital, about 100 babies a day
 - Smaller hospital, about 15 babies a day
- Over the course of the year, which hospital (if either) would probably have more days in which more than 60% of the babies born are male?

how I see things (continued)

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- In order to understand what is going on, we need a model
- For example, we want to figure out whether watching violent movies is correlated with violent behavior
- It is standard to attempt to reject the *null hypothesis* that there is no correlation between the variables
- Consideration
 - data sample
 - experimental design
 - confounding variables
 - choice of method

what we'll cover

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- Probability
- Descriptive statistics
- Inferential Statistics
 - Sampling distributions: confidence intervals, hypothesis tests and p-values
 - Estimating population mean
 - Comparing population means
 - Analysis of variance
 - Univariate and multivariate OLS Regression
 - Analysis of categorical data
 - Data Collection
 - Experimental design

Outline

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- This is SI 544
- I'm Lada Adamic (ladamic@umich.edu)
- Please respond to survey on office hours:
<http://doodle.ch/participation.html?pollId=29v979qxx4tdvmr5>
- Make sure you have access to the cTools site
- Some materials are available at
<http://www-personal.umich.edu/ladamic/courses/si544w08/>

meetings

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

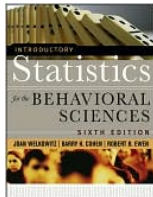
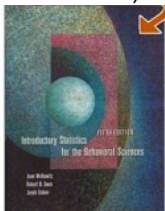
- every Thursday in West Hall 409
- every Tuesday in the DIAD (4th floor of Shapiro library)
 - work on your own laptop
 - work on lab PC or Mac

textbooks

SI 544
Introductory
Statistics and
Data Analysis

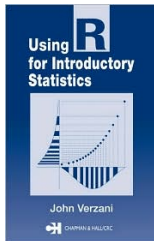
Lada Adamic

- Introductory Statistics for the Behavioral Sciences (5th or 6th Edition) by Welkowitz, Ewen, and Cohen



or

- John Verzani: Using R for Introductory Statistics



R

We will be using R

- open source <http://www.r-project.org/>
- many additional modules available:
`cran.r-project.org/`
- on cTools under Resources you can find several nice online tutorials
- steeper learning curve than most other statistical packages, but ...
 - free — so you can use it whatever your job may be
 - programmable — you can create your own modules
 - you will be able to switch to other software relatively easily

grading scheme

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- 20% midterm (in class)
- 25% final (take home)
- 25% problem sets
 - I will drop your lowest problem set score.
- 20% group project (small)
- 5% news evaluation
- 5% participation
 - attendance
 - speaking up in class
 - posting to cTools

problem sets

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- turn in only in PDF format
- turn in only on cTools
- late assignments
 - can turn in up to 2 days late with a 10% penalty
 - email me for extensions for legit reasons (medical & family emergencies)
include a note about granted extension on cTools when submitting
 - grader Laurel Shipley: email me if there are any questions about grading
- you are encouraged to collaborate with your classmates, but turn in your own work
- run your own R code

cTools participation

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

You can earn participation points by posting content to cTools

- post clarification questions to cTools
There is no point to being stuck on some syntax in R for hours on end. Just ask for help on the cTools forum.
- answer others' questions
- post interesting links

group project

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

The project is not a major project.
At the end you will turn in 3 pages.

- form groups of 3-4 people
- time line
 - Jan. 24 form group and select topic (2 pts)
 - March 4th project progress report (3 pts)
 - April 8 project report (10 pts)
 - April 10 presentations (5 pts)
- examples of topics from last year
 - history of violent arrests and NBA performance
 - temperature and beer consumption
 - number of books and ratings on LibraryThing
 - pacing and rowing performance

news review

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- 5% of your final grade
- find news article and critically compare to original study

Google News

Web Images Video News Maps more »

"study found" Search News Search the Web [Advanced News Search](#)
[Preferences](#)

Results 1 - 10 of

Top Stories

- World
- U.S.
- Business
- Sci/Tech
- Sports
- Entertainment
- Health
- Most Popular


[News Alerts](#)

[RSS](#) | [Atom](#)
[About News](#)

[Mobile News](#)


[About Google News](#)

The **study found** that nicotine levels rose in all types of ...
Boston Globe, United States - Aug 29, 2006
Across the United States, cities and states have moved to ban smoking in all public places, including bars and restaurants. Massachusetts ...

 Study Finds Sharp Drop in the Number of Terrorism Cases Prosecuted
New York Times, United States - 22 hours ago
... Among the most frequent explanations cited by prosecutors, the **study found**, were a lack of evidence of criminal intent by the suspect and "weak or ...
[Study: Terror cases recede](#) Chicago Tribune
[Terror prosecutions fall](#) DetNews.com
[all 293 news articles >](#)

[Bilingual abilities pay off](#)
Statesman Journal, Oregon - 13 hours ago
... in English. In some cases, Hispanics earned \$7,000 more per year than their English-only counterparts, the **study found**. The report ...

[Orange Juice Best at Stopping Kidney Stones](#)
Forbes - 8 hours ago
... The **study found** that orange juice increased levels of citrate in the urine and reduced the crystallization of uric acid and calcium oxalate, the most common ...

 Monday Newspaper Review - Irish Business News and International ...
FinFacts Ireland, Ireland - 17 hours ago
... The **study found** that the mobile industry had succeeded in attracting "early adopters" to the technology, but a "second wave" of consumers was ...

[9 WAYS TO MAKE YOUR SALAD... SUPER](#)
ic Lanarkshire.co.uk, UK - 12 hours ago
... better. An Italian **study found** peppers can cut the risk of cataracts, probably owing to their vitamin C and beta carotene. These ...

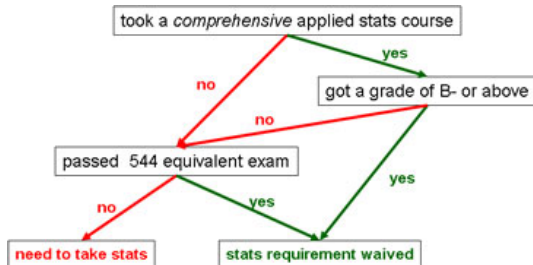
Navigation icons: back, forward, search, etc.

exemptions for HCI and IAR stats requirement

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic



<http://www-personal.umich.edu/~ladamic/courses/si544f06/statswaiver.html>

Courses you can take instead

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- Stat. 350 - Introduction to statistics and data analysis (undergrad, cannot count as cognate)
- Stat. 400 - Applied Statistical Methods
- Stat. 500 - Applied Statistical Methods (note, has 350 as a prerequisite)
- Biostatistics 510: <http://www-personal.umich.edu/~kwelch/510/biostat510.htm>
- Biostatistics 503:
http://www.sph.umich.edu/iscr/caid/display_course.cfm?CourseID=BIOSTAT503
- Biostatistics 553
http://www.sph.umich.edu/iscr/caid/display_course.cfm?courseID=BIOSTAT553
- Sociology 510: statistics
- OMS 501: Applied Business Statistics

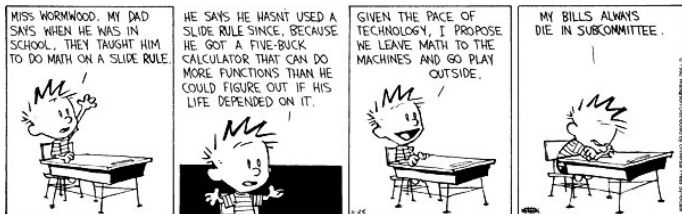
Why you should still take this course

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- neat datasets relevant to HCI & IAR
- we focus on the relevant skills and critical thinking
- we go easy on the math
 - no calculus required
 - some algebra helpful



Outline

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

data types

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- Quantitative
 - Discrete
 - Continuous
- Qualitative
 - Nominal (categorical)
 - Ordinal (rank ordered categories)

exercise: name the data type

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- bacteria count
- occupations of shoppers
- USNWR ranking of university
- marital status
- time (in months) since last auto maintenance
- handedness

Counties with highest rates of kidney cancer

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic



Counties with lowest rates of kidney cancer

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic



summary

SI 544

Introductory
Statistics and
Data Analysis

Lada Adamic

- class logistics (questions?)
- motivation for learning statistics
- starting to gather data!

Next time: R tutorial by Mick McQuaid