DATA 8

Spring 2020

Lecture 38

Conclusion

Announcements

Data Science

Why Data Science

- Unprecedented access to data means that we can make new discoveries and more informed decisions
- Computation is a powerful ally in data processing, visualization, prediction, and statistical inference
- People can agree on evidence and measurement
- Data and computation are everywhere: understanding and interpreting are more important than ever

Limitations of Data Science

- Evidence and measurements are critical ingredients for good decision-making
  - ...but they’re not enough by themselves!
- Data science is a powerful complement to qualitative analysis
  - ...but it’s not a replacement!

How to Analyze Data

Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

Visualize, then quantify!

Perhaps the most important part: Interpretation of the results in the language of the domain, without statistical jargon.

How *Not* to Analyze Data

Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

Visualize, then quantify!

Perhaps the most important part: Interpretation of the results in the language of the domain, without statistical jargon.

How to Analyze Data after Data 8

Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

Visualize, then quantify! Do both using computation.

Perhaps the most important part: Interpretation of the results in the language of the domain, without statistical jargon.

The Design of Data 8

- Table manipulation using Python
- Working with whole distributions, not just means
- Decisions based on sampling: assessing models
- Estimation based on resampling
- Understanding sampling variability
- Prediction

Using Data Science to Understand COVID-19

COVID-19: Remdesivir

Anthony Fauci reports 'quite good news' from Remdesivir trial with control group

By Eric Ting, SFGATE Updated 10:36 am PDT, Wednesday, April 29, 2020

Fauci stated that researchers were primarily looking to evaluate if the drug increased rate of recovery, and found that patients who received the drug recovered in 11 days, and patients that did not receive the drug recovered in 15 days. That 31 percent improvement was deemed "very important" by Fauci.

In addition, the group that received Remdesivir saw a mortality rate of eight percent, compared to a mortality rate of 11 percent in the control group. Fauci stated the difference was not statistically significant, but was still encouraged by the rate of recovery numbers.

According to Fauci, the trial of over 1,000 people was the “first truly high-powered randomized placebo-controlled trial," and shows “a drug can block this virus."

https://www.sfgate.com/coronavirus/article/Remdesivir-control-group-trial-results-Fauci-NIAID-15234422.php

COVID-19: Remdesivir

An independent data and safety monitoring board (DSMB) overseeing the trial met on April 27 to review data and shared their interim analysis with the study team. Based upon their review of the data, they noted that remdesivir was better than placebo from the perspective of the primary endpoint, time to recovery, a metric often used in influenza trials. Recovery in this study was defined as being well enough for hospital discharge or returning to normal activity level.

Preliminary results indicate that patients who received remdesivir had a 31% faster time to recovery than those who received placebo ($p<0.001$). Specifically, the median time to recovery was 11 days for patients treated with remdesivir compared with 15 days for those who received placebo. Results also suggested a survival benefit, with a mortality rate of 8.0% for the group receiving remdesivir versus 11.6% for the placebo group ($p=0.059$).

https://www.niaid.nih.gov/news-events/nih-clinical-trial-shows-remdesivir-accelerates-recovery-advanced-covid-19

COVID-19: Santa Clara County

Santa Clara County COVID-19 cases could be 50 to 85 times higher than reported, Stanford study finds

The study estimates that between 2.49 and 4.16% of people in the county have been infected, according to study co-lead Eran Bendavid, an associate professor of medicine. Bendavid said Santa Clara County has not reached herd immunity, which Gov. Gavin Newsom said in a press conference on Wednesday would be necessary for the state to return to normal.

https://www.stanforddaily.com/2020/04/17/santa-clara-county-covid-19-cases-could-be-50-to-85-times-higher-than-reported-stanford-study-finds/

COVID-19: Santa Clara County

Will Fithian
@wfithian

The authors said by email that they used a built-in Stata function and aren't sure themselves how the software used the input weights. I suspect they misapplied that function (too complicated to tweet why) but I don't know Stata well enough to be sure; it seems neither do they.

1:15 PM · Apr 21, 2020 · Twitter Web App

COVID-19: Santa Clara County

Will Fithian
@wfithian

If we change 369/371 to 366/371, then by my calculations the author's delta method confidence interval for prevalence (based off their "cluster-adjusted interval" for q) becomes [-0.4%, 3.6%].

COVID-19: Santa Clara County

Will Fithian
@wfithian

The errors are not debatable and can be seen in these two screenshots of the supplement: 0.0034, the standard error meant to measure uncertainty about prevalence pi, is not the square root of 0.039, and the variance of a binomial estimate of proportion depends on the sample size.

1:15 PM · Apr 21, 2020 · Twitter Web App

In the main text, we explore three alternative assumptions about the sampling error in sensitivity and specificity. For each of these scenarios, we estimate the variance in sensitivity and specificity by the standard formulas for a binomial outcome: $Var(r) = \hat{r}(1-\hat{r})$ and $Var(s) = \hat{s}(1-\hat{s})$, where $\hat{r}$ and $\hat{s}$ are the estimated values of sensitivity and specificity that pertain to each scenario.

Variance of a sample proportion

COVID-19: Santa Clara County

Sometimes, confidence intervals can be difficult to calculate, and the 95% confidence intervals reported in the study by Bendavid et al. use a standard statistical approximation that often gives accurate results. However, due to the small number of tests used to validate the accuracy of the serology tests, and the nonlinear nature of the analysis, this approximation is inaccurate for the discussed results. The originally reported 95% confidence intervals do not include 0, meaning that they report that if no one in the county had antibodies for SARS-CoV-2, the chance of getting their results are less than 5%. However, in contrast, we show that this probability is actually around 10.9%.

https://medium.com/@steve.yadlowsky/large-variability-in-serology-estimates-of-prevalence-of-sars-cov-2-28151ae4709e

COVID-19: Nicotine and causality

French researchers to test nicotine patches on coronavirus patients

However, the researchers insisted they were not encouraging the population to take up smoking, which carries other potentially fatal health risks and kills 50% of those who take it up. While nicotine may protect those from the virus, smokers who have caught it often develop more serious symptoms because of the toxic effect of tobacco smoke on the lungs, they say.

The team at Pitié-Salpêtrière hospital questioned 480 patients who tested positive for the virus, 350 of whom were hospitalised while the rest with less serious symptoms were allowed home.

It found that of those admitted to hospital, whose median age was 65, only 4.4% were regular smokers. Among those released home, with a median age of 44, 5.3% smoked.

https://www.theguardian.com/world/2020/apr/22/french-study-suggests-smokers-at-lower-risk-of-getting-coronavirus

A Request

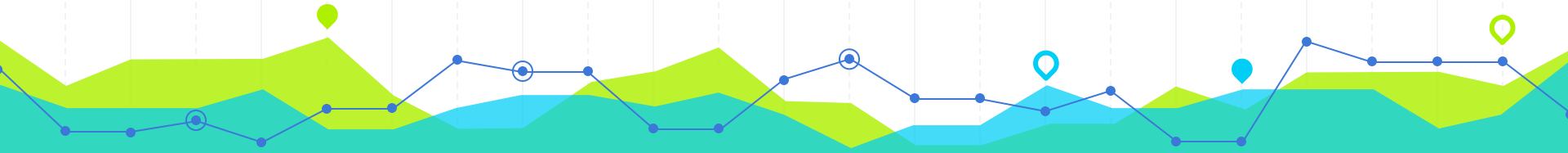
Please fill out the course evaluations!

Course Overview Slides

What's Next?

Guest Presentations

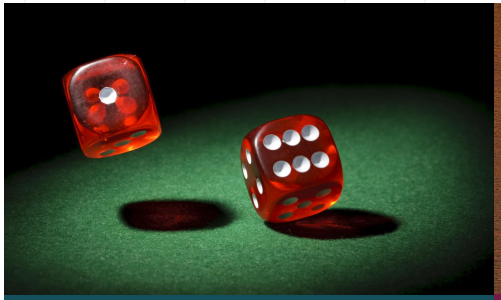
Spring 2020
Student Opportunities



Berkeley Division of Data Science and Information

data.berkeley.edu

Fall 2020 Connectors



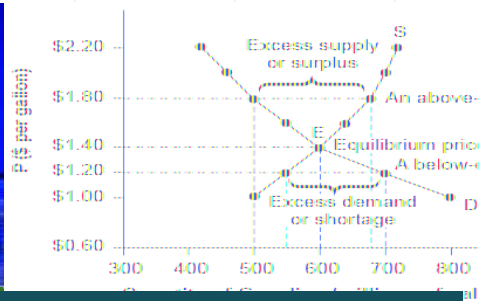
STAT 88
Prob and Stats in Data Science

UGBA 88
Data and Decision

EPS 88
PyEarth: A Python Introduction to Earth Science

COMPSCI 88
Computational Structures

PHYSICS 88
Data Science Applications in Physics



DATA 88-3
Economic Models

Data Scholars

For students who are

- Low income
- First generation college
- Historically underrepresented

Data Scholars serves these populations to support diversity in the Data Science student community all the way through.

Foundations

Concurrent to Data 8

Pathways

Further technical skill & career development

Discovery

Support for research experiences

DSEP Student Teams

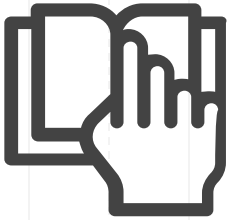
Hone your skills as an educator and data scientist by working with the Data Science Education Program

Help fellow undergrads with data research, academic work, and data science technology.

Peer Consulting

Help deliver Data 8 instruction via online cloud based instruction

Online Learning



Create a national community of practice for institutions to work with and learn from each other.

External Pedagogy

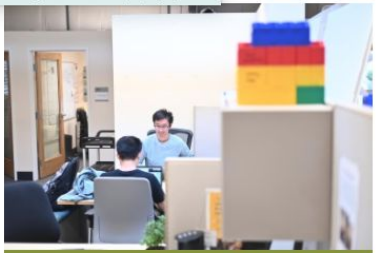
DSEP Student Teams

Hone your skills as an educator and data scientist by working with the Data Science Education Program

Program Support Teams

CDSS div-org-February 2020 (2).pptx

Discovery Operations



Human Resources and Management

Communications Team

Strategic Operations

DSEP Student Teams

Hone your skills as an educator and data scientist by working with the Data Science Education Program

Connector Assistants

Help instructors of Data Science Connector courses deliver and teach material.

Modules

Create curriculum materials for Connectors, Data-Enabled Courses, or short explorations into DS (modules).

Human Context and Ethics

Integrate critical thinking about ethical issues in relation to technology into the Berkeley data science program and community.

Discovery Student Researchers

*Be a **student researcher** in a program that connects students with hands-on data science research- non-profits, start-ups, institutions, etc. Students from underrepresented minority groups and first-time researchers receive priority.*

UCSF University of California San Francisco

Space Sciences Laboratory, University of California Berkeley

BIDS Berkeley Institute for Data Science

The Charles and Louise Travers Department of Political Science, University of California, Berkeley

berkeley urap undergraduate research apprentice program

College of Natural Resources

Digital Humanities at Berkeley

UC Berkeley Center for Community Innovation

Helen Wills Neuroscience Institute

CITRIS Foundry

SKYDECK

https://data.berkeley.edu/research/discovery-program-home

30

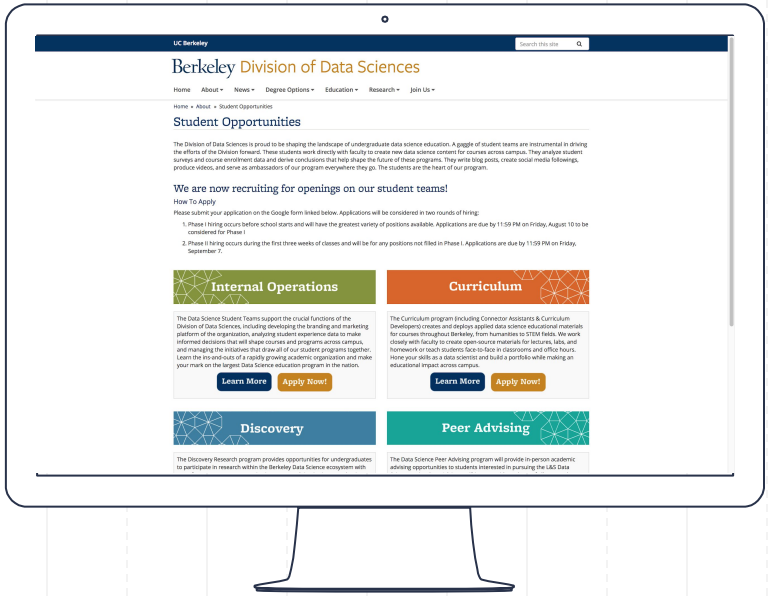
Data Science Peer Advising

Peer Advisors share their diverse knowledge of and experiences with:

- major courses
- different data science major domain emphases
- extracurriculars and student groups on campus
- research opportunities
- and various campus resources.

Peer Advisors work closely with the Data Science Advising Team and gain exposure to the ins-and-outs of a new and increasingly popular interdisciplinary major.

How to Apply



For more information, check out our website at

Applications will re-open in July

https://data.berkeley.edu/academics/resources/student-opportunities

The Team

Staff

- GSIs
- Tutors
- Lab Assistants

Thank you!