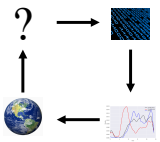
Data Science 100
Lec 3: Problem formulation and experimental design

Slides by:

Bin Yu

binyu@stat.berkeley.edu



Motivations for my DS projects

- External force (Bell Labs)
- Curiosity (UCB neuroscience)
- Social responsibility (UCB remote sensing)
- Intellectual responsibility (UCB precision medicine)
- Opportunity (Bell Labs, UCB developmental biology)

Yu Group

http://www.stat.berkeley.edu/~yugroup/

Statistics, machine learning, causal inference, collaborative research in neuroscience, genomics, and precision medicine.

Solving data problems via statistical and machine learning methods, domain knowledge, and theory, while embedding students/postdocs in labs.

Science projects

- "Mind-reading" (neuroscience)
- Understand V4 neurons via deep learning (neuroscience)
- "Mapping a cell's destiny" (developmental biology)
- Predicting enhancer status with NGS data (genomics)

Presented clip

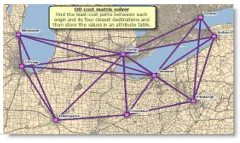
Clip reconstructed from brain activity

Projects from climate science, medicine, and political science

- "Cloud mask" over polar regions as input to climate models (remote sensing)
- How to do organ matching to reduce rejection risk? (precision medicine)
- How to predict a TV ad's partisanship or tone? (political science)
- How to summarize NYT's articles on Russia? (international relations)

Projects from IT industry

- Where is internet traffic going? (network analysis)
- How to compress mixed audio signals to transmit over internet for musicians to play together? (Signal processing)
- How should eBay decide on their web interface? (AB testing)
- How to debug a code? (CS systems)



I. Data Analytics QPR-V

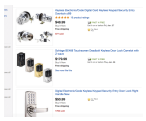
- **Q**uestion 
- **P**opulation 
- **R**epresentative data collection (data neutral) 
- **-** (to be filled in throughout the course)
- **V**etting or validation of answers 

Q ? Question, question, question

- Domain question to answer
- Examples
  - Why didn't the polls work well? 
  - Does smoking cause cancer? 
  - Does BrainPlus IQ work? 

Write down question as record keeping

More questions:

- Which restaurant to go to in Berkeley? 
- Should eBay update their door lock ranking? 
- How to "read mind"? Or how to reconstruct a movie based on brain fMRI signals?   

Two types of questions

- Hypothesis driven: whether a new drug works 
- Discovery-driven: search for new diseases that an existing drug could treat 
- Separate hypothesis generation and decision on hypothesis: discovery phase vs. validation phase

Sample-split is recommended: 

Validation: what makes sense below?

Feasibility of question

- Is the question feasible to answer using data?
- We need to translate the questions into a more precise question:
  Why didn't polls work? → how did we predict the result of the 2016 election in Oct. 2016? → how did we predict the popular vote?
- Did Gallup Poll have the resources (energy, expertise, relevant data) to do the prediction?

P. "Population" in the question

- Population is the relevant group of people (objects, units) that the data-driven answer to the question will be applied to

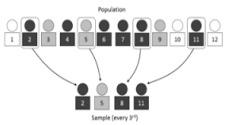

- Write the population down as a record for any DS project

"Population" for the 2016 election

- Oracle (future) population
  all the votes casted on election day
- Short of magic, we want to predict or guess these votes ahead of time
- How to go about this?

R. Representative data collection

- Data to answer the question should be "representative" of the relevant group (or population), or data neutral



Never a brand new problem…

- There is always prior knowledge or data
- Qualitative: literature, human co-workers
- Quantitative: previous data, new pilot study

Election example: data collection

- Population: all the votes to be casted on election day in CA -- impossible since we can't time travel
  Even if we can, too much money to ask everyone
- How about a representative sample of the election day votes, also for the sake of money
  -- can't time travel so possible only under assumptions

Two layers of uncertainty on Nov. 8

- Who are going to vote on election day?
- How is a voter going to vote?
- We can't ask all voters, even if we could, people change their minds – for 2016, about 12% undecided voters, compared with 4-5% in the past

Assumptions in the election problem

1. Today's votes are the same as on the election day: People are asked what they would vote if they are voting today –

2. People are telling the truth

3. Undecided voters vote similarly as decided voters

4. The polled group is representative of the voter population

How to make the sample representative or data neutral?

- A whole field of statistics, **Survey Sampling**, has been devoted to this. It recommends random sampling of different kinds. (Take **Stat 152**)
- Simplest: **simple random sampling (SRS)** without (or with) replacement – putting all entities of the population in a hat and randomly drawing one by one

One particular poll: Gallup Poll

- Simplest form: 1000 SRS samples from numbers of phones (landline and cell phones) (one landline can map to many voters).
- What is the population? Is it the same as the voter population?

- What if people without phones voted very differently from people with phones?

Gallup Poll: some quick calculations

- Suppose Gallup poll is an SRS of the voter population on election day:

 =  , and previous assumptions

139 million votes, 68 for Trump, 71 for Clinton – less than 2% diff.

- Margin of error is about 3%, one Gallup Poll is not enough to predict such a close race **even if all the assumptions hold and they do not.**

Vetting results: election

- Votes on election day
- Prediction on popular votes held (by luck), but not for election outcome.

Vetting data results in general

- Prediction on test data
- Stability analysis
- Post-modeling EDA or visualization
- Domain knowledge verification
- …

Longer time scale

- Down-stream consequences
- Further studies
- …

Recall "danger zone"

- Danger zone: algorithms applied to domain problems without understanding of statistical concepts/issues such as population, representative data collection, and uncertainty, …
- For election and personalized medicine, there is NO averaging effect to aid algorithms as in IT applications where ML algorithms have been traditionally successful.

One has to get representative samples AND try to use other information to reduce variability for a one-time shot!

Clinton Campaign fell into danger zone?

POLITICO

2016

How Clinton lost Michigan — and blew the election

Across battlegrounds, Democrats blame HQ's stubborn commitment to a one-size-fits-all strategy.

By EDWARD-ISAAC DOVERE | 12/14/16 05:08 AM EST

http://www.politico.com/story/2016/12/michigan-hillary-clinton-trump-232547

"In results that narrow, Clinton's loss could be attributed to any number of factors — FBI Director Jim Comey's letter shifting late deciders, the lack of a compelling economic message, the apparent Russian hacking. But heartbroken and frustrated in-state battleground operatives worry that a lesson being missed is a simple one: Get the basics of campaigning right."

"Clinton never even stopped by a United Auto Workers union hall in Michigan, though a person involved with the campaign noted bitterly that the UAW flaked on GOTV commitments in the final days, and that AFSCME never even made any, despite months of appeals."

Thanks to J. Sekhon

2016

How Clinton lost Michigan — and blew the election

Across battlegrounds, Democrats blame HQ's stubborn commitment to a one-size-fits-all strategy.

By EDWARD-ISAAC DOVERE | 12/14/16 05:08 AM EST

"Brooklyn mandated that they not worry about data entry. Operatives watched packets of real-time voter information piled up in bins at the coordinated campaign headquarters. The sheets were updated only when they got ripped, or soaked with coffee. Existing packets with notes from the volunteers, including highlighting how much Trump inclination there was among some of the white male union members the Clinton campaign was sure would be with her, were tossed in the garbage."

"The Brooklyn command believed that television and limited direct mail and digital efforts were the only way to win over voters, people familiar with the thinking at headquarters said. Guided by polls that showed the Midwestern states safer, the campaign spent, according to one internal estimate, about 3 percent as much in Michigan and Wisconsin as it spent in Florida, Ohio and North Carolina. Most voters in Michigan didn't see a television ad until the final week."

"Most importantly, multiple operatives said, the Clinton campaign dismissed what's known as in-person "persuasion" — no one was knocking on doors trying to drum up support for the Democratic nominee, which also meant no one was hearing directly from voters aside from voters they'd already assumed were likely Clinton voters, no one tracking how feelings about the race and the candidates were evolving. This left no information to check the polling models against — which might have, for example, showed the campaign that some of the white male union members they had expected to be likely Clinton voters actually veering toward Trump — and no early warning system that the race was turning against them in ways that their daily tracking polls weren't picking up."

Data science lessons

- Analytical algorithms CAN NOT automatically detect non-representative or biased samples
- Shoe leather work needs to be taken seriously and with analytical algorithms:

Information about undecided voters and people who do not respond to polls can be obtained only through ground operatives in their talking and interactions with such people.

30 sec. meditation …

II. Experimental design

The science and subfield of statistics about how to collect data effectively…

R. A. Fisher (1890-1962)
Founding father of Modern Statistics
Geneticist

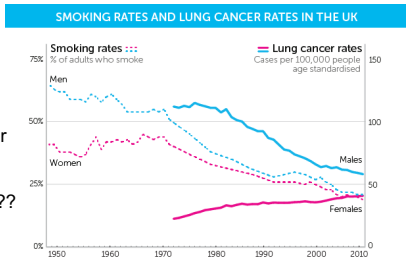
Smoking causes lung cancer?

Population?

Thanks to R. Barter for help on some smoking-cancer slides

Does smoking cause cancer?

UK Cancer Research



Can we conclude that

smoking causes cancer for men;

but is good for women??

Possible explanations based on "colleague-sourcing"

- Lung cancer is not all the same
- Only 10-15% is closely associated to smoking
- Another type is not, but occurs more among women
- Women started smoking later than men and the cohort with peak smoking is still working its way through the population
- Women are more susceptible to tobacco toxins
- …

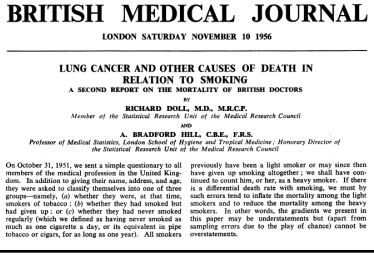
Thanks to P. Stark, P. Ding, J. Sekhon

Lessons

- Ecological correlation does not imply correlation at person level (ecological correlation = correlation between rates)
- Association (correlation) is not causation
- Confounding factors are always lurking in the back

(confounding factor: a possible driver for both smoking and lung cancer, e.g. genetics)

The first solid epidemiological evidence, or observational study



BRITISH MEDICAL JOURNAL

LONDON SATURDAY NOVEMBER 10 1956

LUNG CANCER AND OTHER CAUSES OF DEATH IN RELATION TO SMOKING

A SECOND REPORT ON THE MORTALITY OF BRITISH DOCTORS

BY

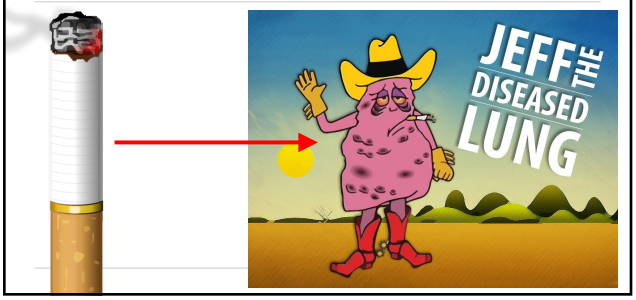
RICHARD DOLL, M.D., M.R.C.P.
Member of the Statistical Research Unit of the Medical Research Council

AND

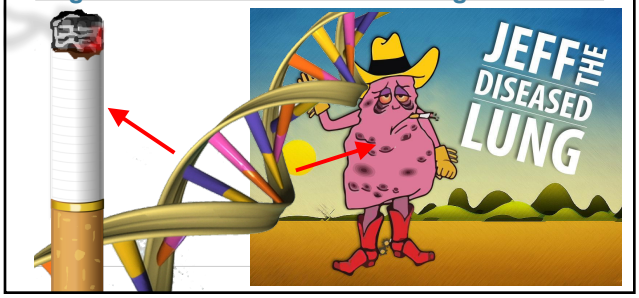
A. BRADFORD HILL, C.B.E., F.R.S.
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

On October 31, 1951, we sent a simple questionary to all members of the medical profession in the United Kingdom. In addition to giving their name, address, and age, they were asked to classify themselves into one of three groups—namely, (a) whether they were, at that time, smokers of tobacco; (b) whether they had smoked but had given up; or (c) whether they had never smoked regularly (which we defined as having never smoked as much as one cigarette a day, or its equivalent in pipe tobacco or cigars, for as long as one year). All smokers previously have been a light smoker or may since then have given up smoking altogether; we shall have continued to count him, or her, as a heavy smoker. If there is a differential death rate with smoking, we must by such errors tend to inflate the mortality among the light smokers and to reduce the mortality among the heavy smokers. In other words, the gradients we present in this paper may be understatements but (apart from sampling errors due to the play of chance) cannot be overstatements.

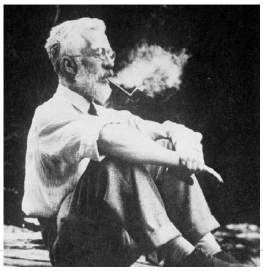
Cigarettes cause lung cancer!



Genetics cause both smoking and lung cancer? Or genetics could be a confounding factor

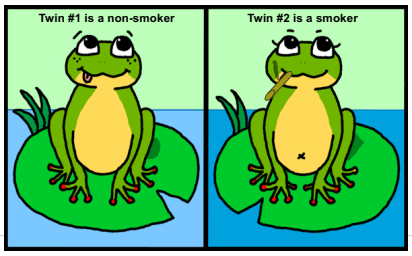


R. A. Fisher strongly believed in a common genetic cause!



17 February 1890 –
29 July 1962

Control for genetic differences: compare identical twins!



The evidence kept piling in from **observational studies**… Until finally in 1964

SMOKING *and* HEALTH

REPORT OF THE ADVISORY COMMITTEE
TO THE SURGEON GENERAL
OF THE PUBLIC HEALTH SERVICE

"Since 1939 there have been 29 retrospective studies of lung cancer alone which have varying degrees of completeness and validity."

"After appraising 16 independent studies carried on in five countries over a period of 18 years, this group concluded that there is a causal relationship between excessive smoking of cigarettes and lung cancer."

U.S DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

What is an ideal experiment to see whether smoking causes cancer?

- What is the population?
- Are you going to ask some people to smoke? What is the difference from people choose to smoke?
- And then what?

Does BrainPlus IQ work?

Anderson Cooper: Stephen Hawking Predicts, "This Pill Will Change Humanity" And It's What I Credit My $20 Million Net Worth To

Featured in: YAHOO! GQ WebMD TIME People Aol.

"This pill unlocks your brain power, allowing you to adventure further inside your own brain than ever before. This is the most groundbreaking BrainPlus IQ ever created, and we had to showcase it to the world!"

National Geographic
Limited Edition Cover Page

How do we translate "change humanity" into a measureable outcome?

How do we measure "adventure further inside your own brain"?

Recently Hawking made some comments in an interview with Anderson Cooper about BrainPlus IQ that would become the biggest event in human history.

Does BrainPlus IQ work?

- What is the population? All people who want to take BrainPlus IQ (many are unborn yet)…
- An easier question: is BrainPlus IQ better on average for people who have signed up for a study conducted by the company?
  Translate "work" into IQ measure…

Population: two potential outcomes (IQ test score) while taking BrainPlus IQ and taking a placebo, respectively

Data collection by observing the population

- Administer BrainPlus IQ on Monday
- Administer Placebo on Tuesday

What assumption are we making?

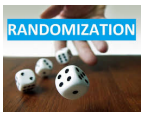
If want to remove the assumption

- Administer BrainPlus IQ to half of the group and placebo to the other half

How would you choose the half?

Gold Standard of Causal Evidence

Randomization



which often relies on pseduo-random number generators (PRNGs) – deterministic processes that can be flawed …

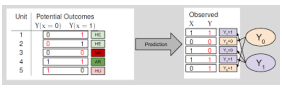
It randomly assigns each subject to the treatment or control group – to take BrainPlus IQ or Placebo – or a random split.

Virtues of Randomization

- Reduce or combat confounding (often not to zero)
- Ground probabilistic reasoning and calculation

Neyman-Rubin model

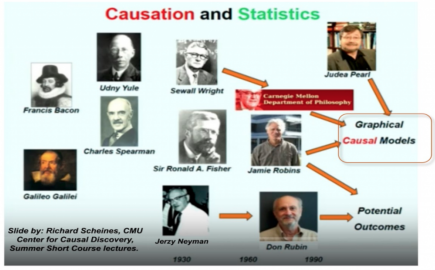
- Population: potential outcomes under treatment and control for the subjects under study



- Super-population: the subjects in the study are representative of the population needing the drug
  -- almost always argued by experts using prior information – judgement call

Neyman was the founder of Berkeley Statistics;
Rubin is a Harvard Statistics Professor.

A bit history…



Three principles of experimental design

- Replication
- **Randomization**
- Blocking (reducing variability by using extra information – highly needed for the election situation)

Summary

- Q (not unique: translating Q in English into a Q about data…)
- P
- R
- -
- V
- Association is not causation
- SRS, Randomization (randomized experiment vs. observational study)
- Confounding factors
- Ecological correlation

Last but not least:

a data-driven claim

Marriage "causes" drowning...