STA 305/1004 - Design of Scientific Studies

Sections L0101 & L0201, Winter 2020

Dr. Shivon Sue-Chee

UNIVERSITY OF TORONTO

January 6-7, 2020

Course Instructor Info

- **Dr. Shivon Sue-Chee**
  - Office: Health Science Building, Office 384
  - Office hours:
    - Tuesdays, 10:10-12:10 in HS 384, from Jan 14
    - *or by appointment*
  - Email: shivon.sue.chee@utoronto.ca

- Course Website:

  `https://q.utoronto.ca/`

What is this course about?

Objective: Understand the ideas, principles, and considerations that are common to the design and analysis of scientific studies:

- Experiments, Observational studies
- Selection bias, Casual inference
- Mathematical statistics for experimental design
- Power and sample size calculations
- Block, Factorial and Split-plot designs

TEXTBOOK

`http://utstat.toronto.edu/~nathan/designscistudynotes.htm`

Statistical Computing:

- *R* (and *RStudio*)
- Support given by Instructor and TAs

Who can take this course?

- Undergrads and Grads with STA302/1001 or equivalent preparation
- Notes:
  - Pre-requisites are strictly enforced
  - No waiver forms accepted due to departmental policy
  - Email Gillis (gillis.aning@utoronto.ca) if you deferred the STA302 exam or have a transfer credit, to make sure you won't be removed.

How will you be evaluated?

	STA305	**STA1004**	Date	Time
Assignment 1	5%	10%	Feb. 4	due by 10pm
Assignment 2	10%	15%	Mar. 30	due by 10pm
Test (L0101)	35%	30%	Feb. 26	11:10-12:40
Test (L0201)	35%	30%	Feb. 27	15:10-16:40
Final Exam	50%	45%	In April	(3 hours)

Grads and undergrads will be evaluated based on different schemes.

What is "BIG DATA"?

- Defined by 3V's:
  1. Volume
  2. Velocity
  3. Variety
- " *big data may be as important to business - and society - as the Internet has become. Why? More data lead to more accurate analyses.*"
  (SAS, `http://www.sas.com/en_id/insights/big-data/what-is-bigdata.html`)
- Is statistical sampling and randomization still relevant in the era of Big Data? (Xiao-Li Meng)

What is "BIG DATA"?

- Defined by 3V's:
  1. Volume
  2. Velocity
  3. Variety
- " *big data may be as important to business - and society - as the Internet has become. Why? More data lead to more accurate analyses.*"
  (SAS, `http://www.sas.com/en_id/insights/big-data/what-is-bigdata.html`)
- Is statistical sampling and randomization still relevant in the era of Big Data? (Xiao-Li Meng)

BIG DATA

Q: In 2015 the population of Canada was 35.8 million people.

To estimate the mean number of hours spent on the Internet is it better to:

(A) take a simple random sample of 100 people (and ask about hours spent on internet) and estimate the mean number of hours spent on the Internet; or

(B) use a large database (e.g., millions of people) that contain hours spent on the Internet for each person?

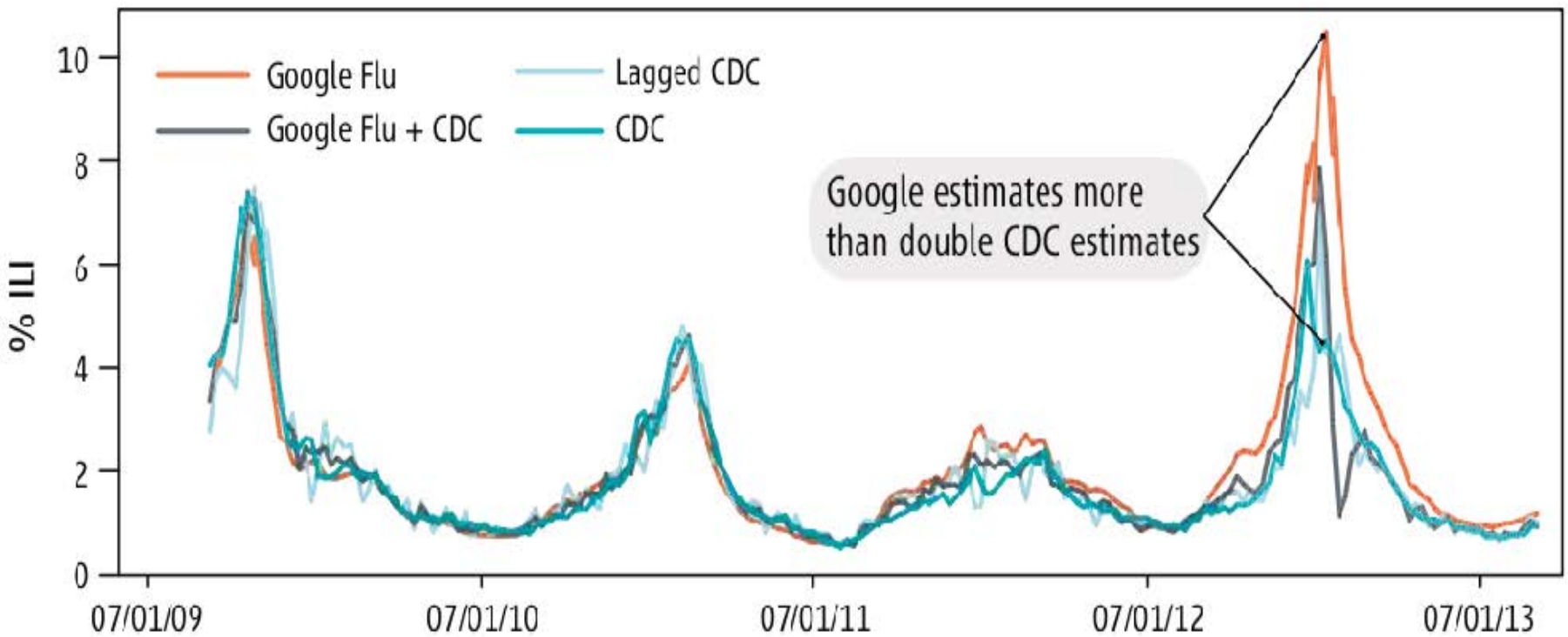
BIG DATA

- To have equivalent precision of a random sample of 100 people a database would have to contain over 96% of the population, i.e., 34.3 million people.
- This illustrates the power of random sampling and the danger of putting faith in "Big Data" simply because it's big.

BIG DATA

- Most "big data" is not obtained from instruments designed to produce valid and reliable data amenable for scientific analysis.

  Eg, Google Flu (Lazer et al., Science 14 March 2014)



SCIENTIFIC PROCESS

- *Box, Hunter and Hunter (2nd ed.), Figure 1.1*

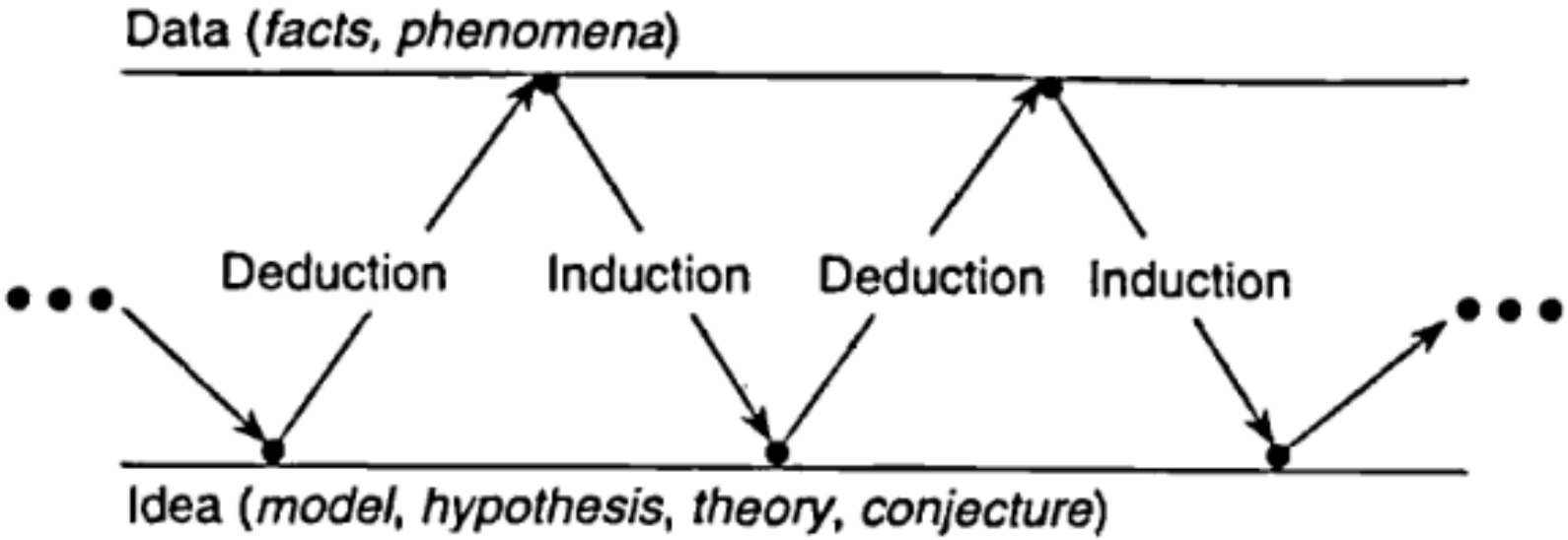


Figure 1.1. Iterative learning process.

WHY DESIGN?

Why should scientific studies be designed?

- Avoid bias
- Variance reduction
- System optimization

WHY DESIGN?

Why should scientific studies be designed?

- Avoid bias
- Variance reduction
- System optimization

Motivating Example: Hotelling's Weighing Problem

- Harold Hotelling in 1949
- **Want to measure the mass of two apples A and B using an old-fashioned two-pan balance scale**
- What is the optimal measurement strategy?

METHODS:

1. Weigh each apple separately
2. Weigh both in one pan and, Weigh one apple in one pan and one apple in the other pan

Motivating Example: Hotelling's Weighing Problem

Apples have unknown weights, w_1, w_2

METHOD 1:

1. Weigh one apple:
2. Weigh the other apple:

METHOD 2:

1. Weigh both in one pan to obtain the *sum* of the two weights:
2. Weigh one apple in one pan and one apple in the other pan to obtain the *difference* between the two weights:

Motivating Example: Hotelling's Weighing Problem

- **Suppose each individual weighing has variance σ^2**
- Observe: METHOD 1 METHOD 2

	Parameter	Estimate	Variance of estimate
Method 1	w_1		
	w_2		
Method 2	w_1		
	w_2		

Motivating Example: Hotelling's Weighing Problem

How many individual weighings (METHOD 1) would be required to achieve the same precision as METHOD 2?

Determining Causation

- A major issue in experimentation is confusion of correlation with causation.

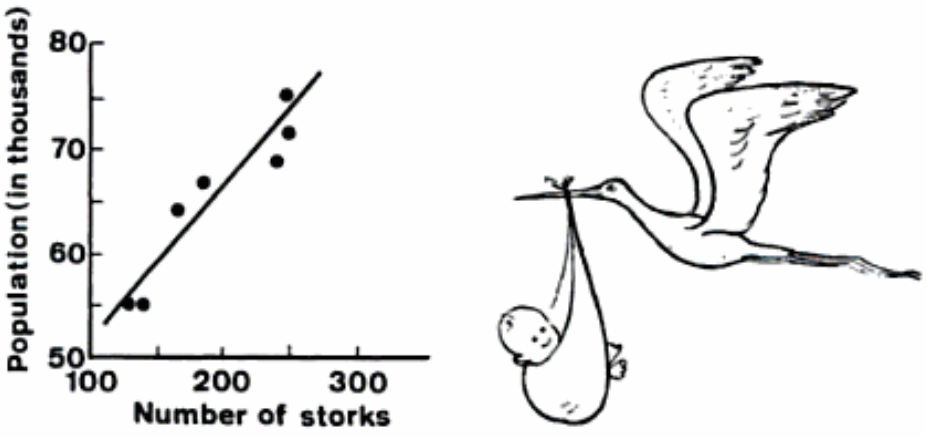
- Randomized Experiments:
  - (pg. 20, Imbens and Rubin, 2015):
    "... the assignment mechanism is under the control of the experimenter, and the probability of any assignment of treatments across the units in the experiment is entirely knowable before the experiment begins."

  - most credible basis for determining cause and effect relationships

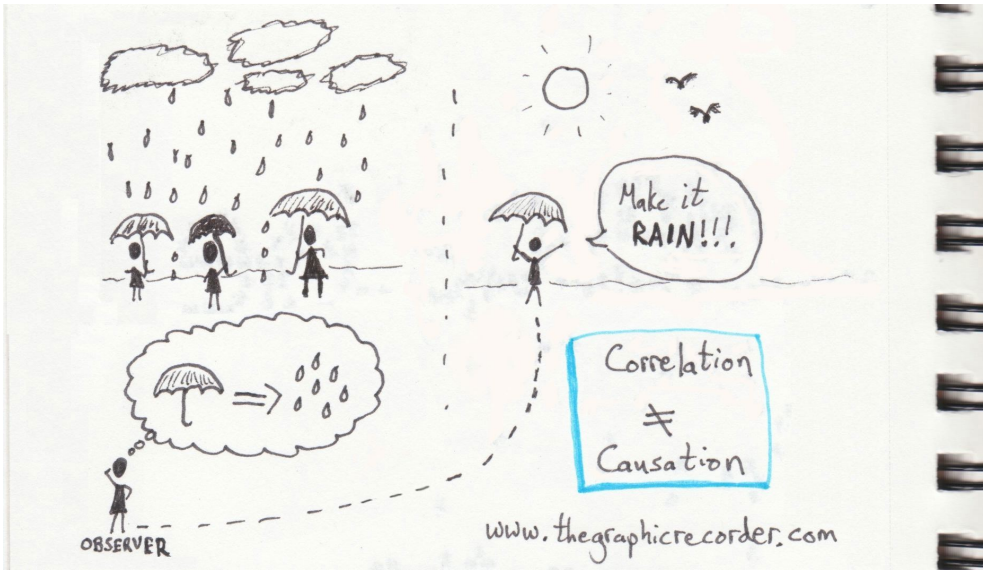
  - relied on by Health Canada, the U.S. Food and Drug Administration, European Medicines Agency, and other regulatory agencies in their drug approval processes

Casual Inference?

Do storks bring babies?



Do umbrellas make it rain?



Where to get help?

- **Don't spin your wheels, ask for help!!**
- Course Website:

  https://q.utoronto.ca/
- Do practice problems/assignments.
- Post questions and review answers on Piazza discussion forum:

  piazza.com/utoronto.ca/winter2020/sta3051004
- Visit the instructor and/or TAs during office hours.
- Email the instructor in cases of emergencies or personal matters.

What's Next?

- Next class: Review of Statistical Theory
- Get R and RStudio