Design of Experiments and Observational Studies

Introduction

STA305: Design and Analysis of Experiments
STA1004: Introductory Experimental Designs

Introduction to Observational and Designed Studies

A General View on a Population

- Objective is to know the features of the population
- Using sample (subset) of the population
- Sample: random/representative
- Collect/generate data
- Draw inference on the population features

Why to Design Scientific Studies when lots of data are available?

- The available data might not be sufficient to answer the questions of interest.

Ways to collect/generate data from a population

1. Observational study (OS)

- **Sample survey:** Sampled units respond to the questions asked by the researcher; aimed to estimate population mean; single variable.

- **OS:** More general is the case of several variables/ measurements on the same population/ sampling units.

- Associations among variables [simple graphical displays] are of interest.

- Sampled units are observed by the researchers than just the respondents. A natural question is:

- Are the relationships found due to direct association, indirect assoc. via other variables, causal relationship or by chance/coincidence?

- Such an area of study is called observational study (OS).

Ways to collect/generate data from a population

2. Designed Experiments (DE)

- The OS provide insight into the relevance of relationships and further exploration in terms of:

  formulation of research hypothesis, detection/testing and estimation of related parameters

  may provide a basis or premises for concluding a cause-and-effect relationship (Medical studies, industrial application, agricultural research, etc.).

- This study area is called Design of Experiments (DE).

- In this lecture, to introduce the concepts and salient features of OS and DE, I will use a few examples.

1. Observational Studies (OS)

- Multiple variables: response variable(s), explanatory variable(s) (may arise with time).

- Associations between/among variables are of interest.

- The variables influence each other.

- The explanatory variables may be associated among themselves.

- Researcher/ experimenter has no role in intervening with the variables association/ relationships.

- In an OS, for a variable considered as response variable, the explanatory variables are not the only variables contributing to the response variable.

Observational Studies (OS) Examples

1. Sampled units from a popn. **Observe the cardiovascular disease risk on whether or not one is a vegan.** Association.

2. Observe "a response or an outcome" variable on units/individuals in say two groups (say male and female); noted a large difference in group means of the measurements; association.

Is "gender" the cause for the difference?

3. Example:The NHEFS (National Health Epidemiologic Follow-up Studies) survey designed to investigate the relationships between clinical, nutritional, and behavioral factors assessed in the first National Health and Nutrition Examination Survey NHANES I.

Smoking showed significant association with the weight gain. However, **this study cannot establish that quit smoking is cause of weight gain.** (Link: https://www.cdc.gov/nhanes/nhefs/nhefs.htm)

4. Example: A study published in the New England Journal of Medicine (May 2012) on the relationship between coffee drinking and mortality. Suppose the coffee drinkers tended to live longer so there's a positive association.

A question to ponder: **Is it coffee that causes long life?** Or are there other possible reasons for the longevity?

2. Design of Experiments (DE)

- In DE, the researcher intervenes with the exp. units via treatment assignments and observes the response.

- Good experiments follow the principles of experimentation, i.e., randomization, replication, and control of error (blocking).

- Response variable = outcome of interest.

- Explanatory variable: treatment or other variables that would have created the variation in the response.

- Principles and their implementation

- Randomized experiment: treatments are assigned to experimental units (participants) on a random basis.

- Use hand-drawn sketches

Design of Experiments (DE): Examples

- **1. Rigorous veggie diet found to slash cholesterol**

- DE: Individuals in a population, randomly assigned the rigorous veggie diet(s) and a control or non-veggie diet.

- Reduction in cholesterol measured and compared.

- Researcher introduced an intervention through (random) assignment of diets to the individuals

- **OS** version: If a population were observed for the diet taken and cholesterol measured,

- The researcher would not have intervened with any assignment of diets to the individuals.

- He/she simply observed the diets and cholesterol levels of the individuals selected; the diets were not assigned to the individuals.

- **2. Blood pressure (BP) reduction**

- New drug vs standard drug

- random assignments/blinding

- **3. Effect of temperature and material type on the battery life (days)**

- DC Montgomery textbook

- **4. Comparing yield productivity of three new wheat varieties and a standard variety in the region**

- Principles/implementation

- Use hand-drawn sketches

OS informative for a DE

- Association between the explanatory variables and intended response variable (OS) leads to formulate a designed study, cause-effect relationship

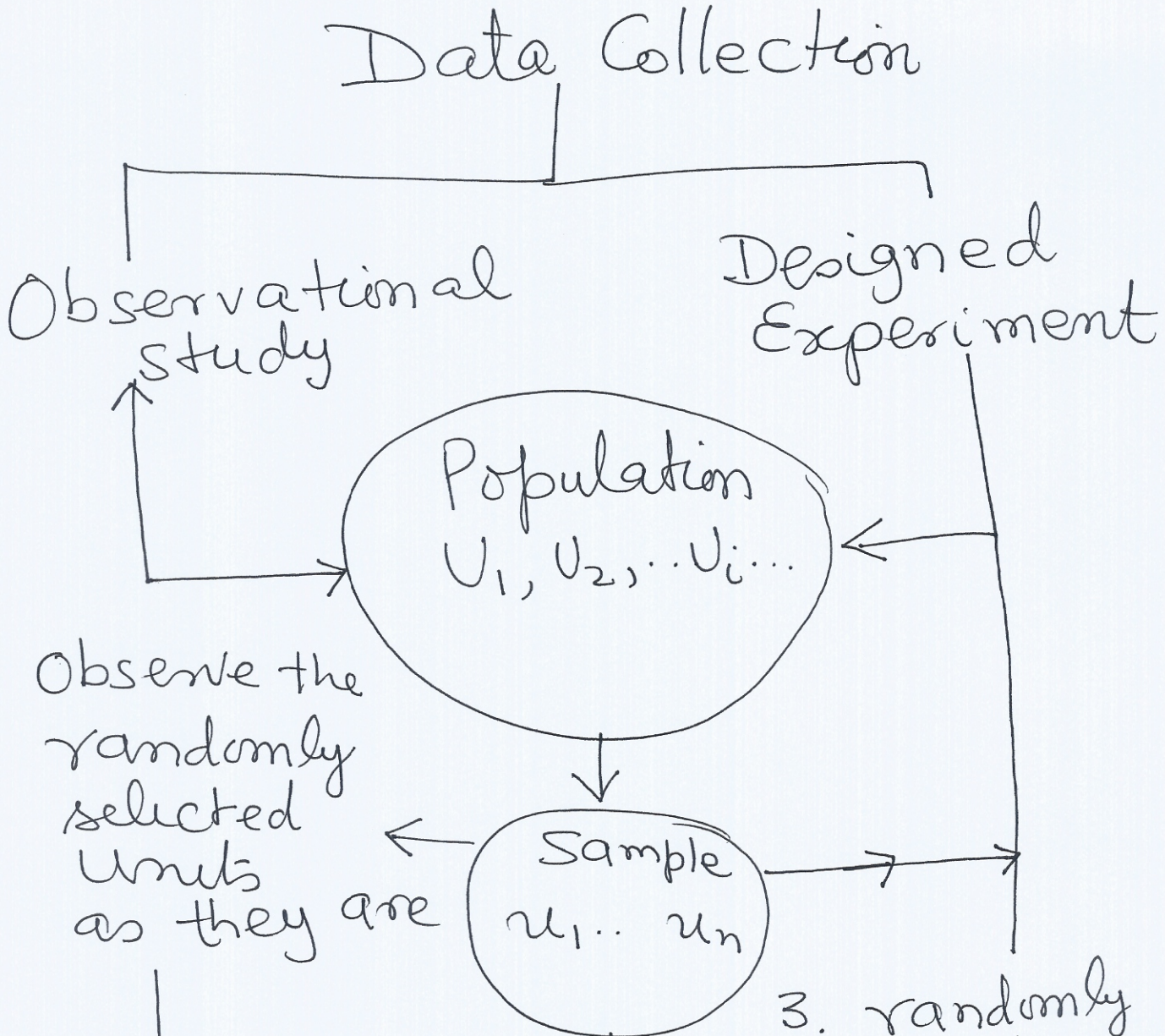
- Identify population/experimental units and formulate the treatment (cause factors)

- Design the study based on the principles (treatment - exp. units assignments)– i.e., example, randomization, replication and reduction of errors (local control)

- Data model assumptions and diagnostics (link response to the treatment effects and exp. unit structures)

- Inferences: unbiased and precise estimates, tests of hypothesis– p-value, and cost

Data Collection

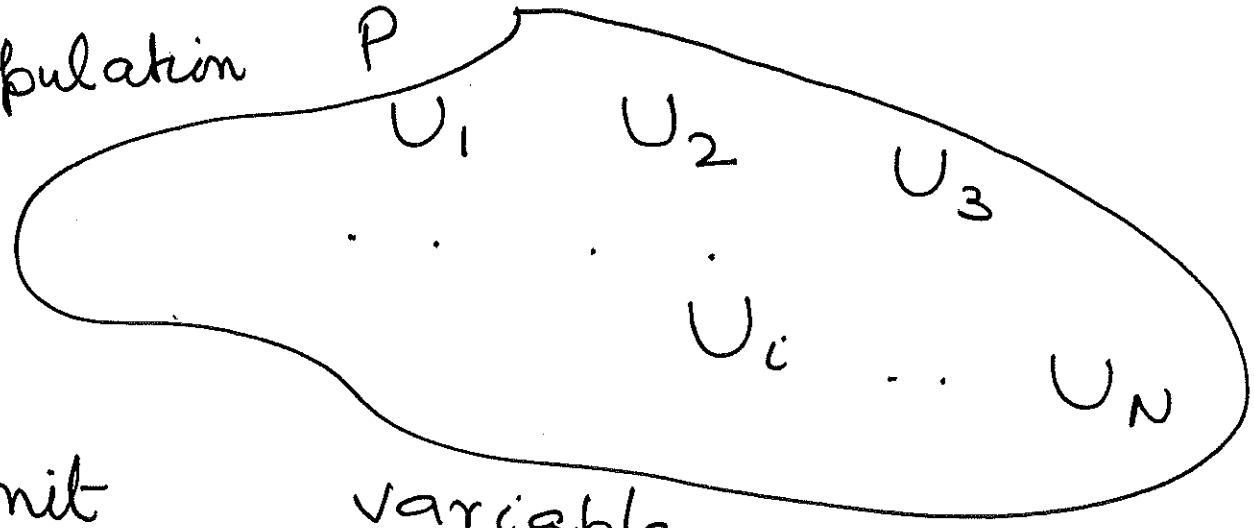


1 Observe the randomly selected units as they are

Study the associations among variables, factors etc

2 Survey – selected units respond to questions of the interviewer.
– Estimate popn mean, etc.

3. randomly assign the treatments to the expt units, i.e. intervention made, observe/measure the response.
– causal effect / association



Unit variables

i, U_i X_1 X_2 ... X_p, Y

observed / measured X_{1i} X_{2i} ... X_{pi}, Y_i

Random sample

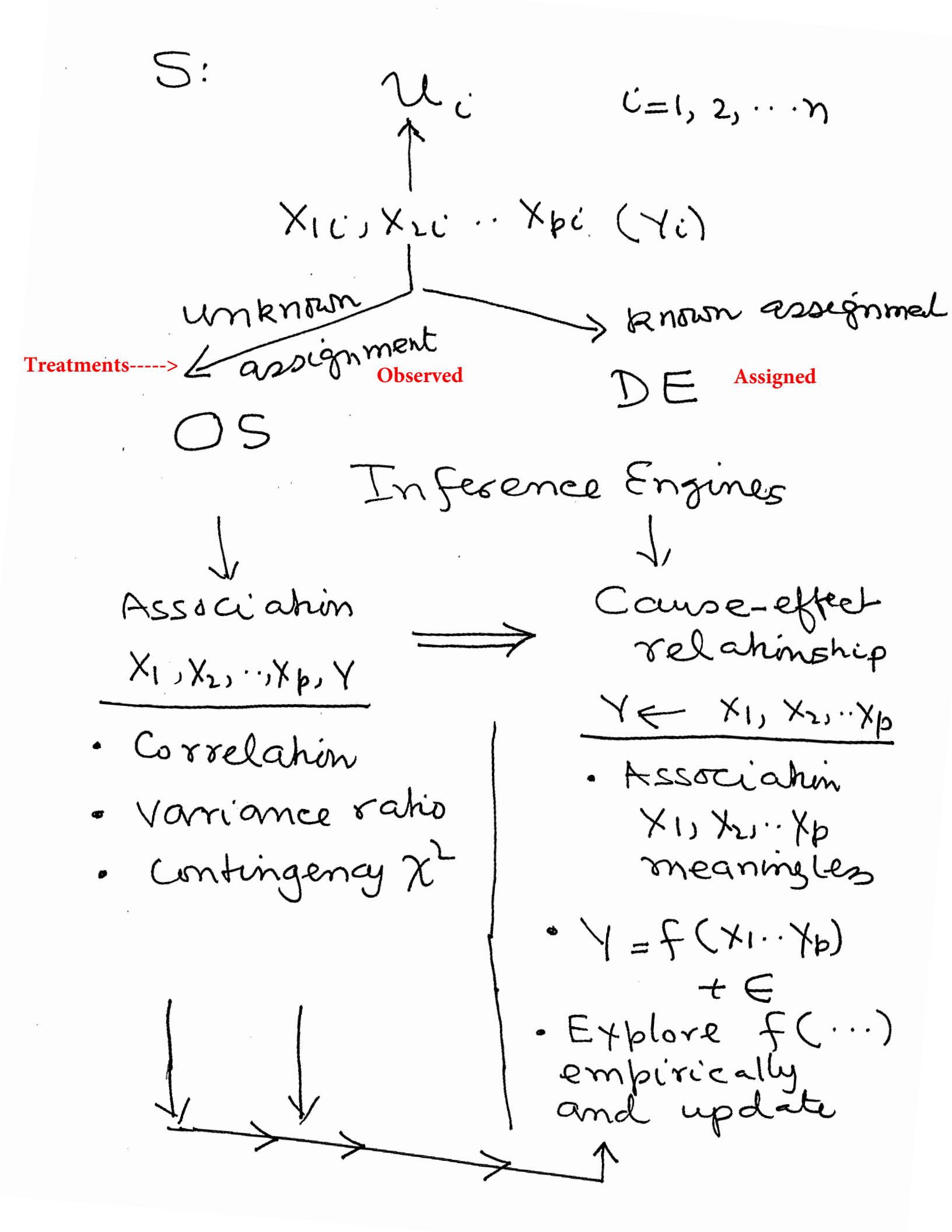
$S: \{u_1, u_2 .. u_n\}$

S a subset of P

Variables:

Explanatory : $X_1, X_2 .. X_p$

Response : Y say.


S:
u_i
$i = 1, 2, \cdots n$
$X_{1i}, X_{2i} \cdots X_{pi}$ (Y_i)
unknown assignment
known assignment
Treatments----->
Observed
Assigned
DE
OS
Inference Engines
Association
$X_1, X_2, \cdots, X_p, Y$
Correlation
Variance ratio
Contingency χ^2
Cause-effect relationship
$Y \leftarrow X_1, X_2, \cdots X_p$
Association
$X_1, X_2, \cdots X_p$
meaningless
$Y = f(X_1 \cdots X_p) + \epsilon$
Explore $f(\cdots)$ empirically and update

Week 1- Lecture 1, 10 September 2021

Acknowledgment

" This document has been prepared by Professor Nathan Taback.

I am grateful to Professor Nathan Taback for providing me this document for presentation and discussion in the class of STA305 Fall 2021.

Murari Singh
"

- WHY DESIGN?
- Why should scientific studies be designed?
- Avoid bias
- Variance reduction
- System optimization

Selected topics

- Big Data
- Professor Wald consultancy – Bullets on fighter planes
- Experimental errors and Professor Hunter's video
- Yate's weighing design
- Spurious correlation
- Regression
- R codes

BIG DATA

"… 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-big-data.html)

Big data

- Bias, precision, accuracy
- Xiao-Li Meng (2014)
- $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\hat{\theta} - \theta))^2$,or

  $MSE(\hat{\theta}) = Variance + Bias^2$
- $\bar{x}_a$ and $\bar{x}_s$
- $MSE(\bar{x}_a) < MSE(\bar{x}_s)$ if
- $f_a = \frac{n_s \rho_N^2}{1 + n_s \rho_N^2}$
- where f_a = is the fraction of the population in the database, ρ_N is the correlation between the response being recorded and the recorded value, and n_s is the size of the random sample. For example, if n_s = 100 the database would need over 96% of the population if ρ=0.5 to guarantee that $MSE(\bar{x}_a) < MSE(\bar{x}_s)$. In our example this would require a database with 34,319,616 Canadians.

BIG DATA

In 2015 the population of Canada is 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

- An equivalent precision of a random sample of 100 people a database would have to contain over 96% of the population 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.

ABRAHAM WALD AND THE MISSING BULLET HOLES

ABRAHAM WALD AND THE MISSING BULLET HOLES

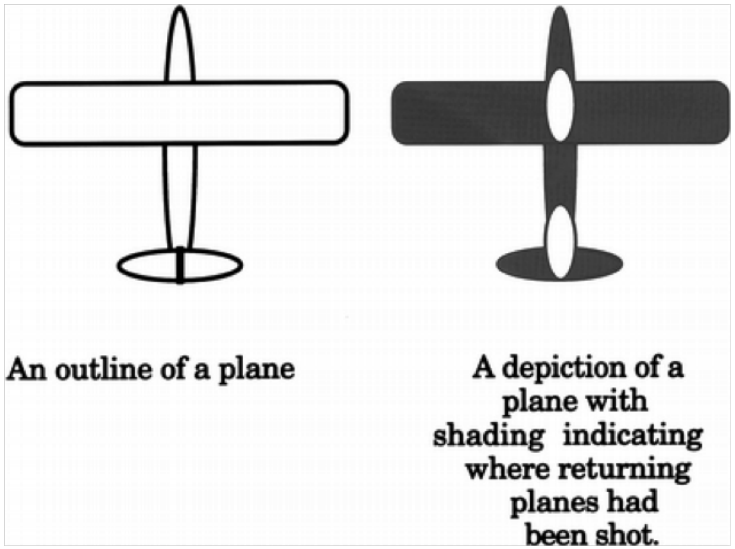
- Abraham Wald born in 1902 in Austria.
- Emigrated to the U.S. and eventually became a professor at Columbia.

See:
http://www.ams.org/publicoutreach/feature-column/fc-2016-06

ABRAHAM WALD AND THE MISSING BULLET HOLES

Question: You don't want planes to get shot down by enemy fighters, so you amour them. But armour makes planes heavier, and are less maneuverable and use more fuel. Armouring planes too much is a problem; armouring the planes too little is a problem.

Somewhere in between there's an optimum.



ABRAHAM WALD AND THE MISSING BULLET HOLES

Planes were covered in bullet holes, but the holes weren't uniformly distributed across the aircraft.

ABRAHAM WALD AND THE MISSING BULLET HOLES

Data from American planes that came back from engagements over Europe.

What parts of the plane has the greatest need for armour?

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

ABRAHAM WALD AND THE MISSING BULLET HOLES

The officers saw an opportunity for efficiency.

Get the same protection with less armour if you concentrate on places with the greatest need.

They asked Wald how much more armour belonged on those parts of the plane.

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Professor J. S. Hunter's presentation on:

What Is Design of Experiments? Part 1

https://www.youtube.com/watch?v=NoVlRAq0Uxs

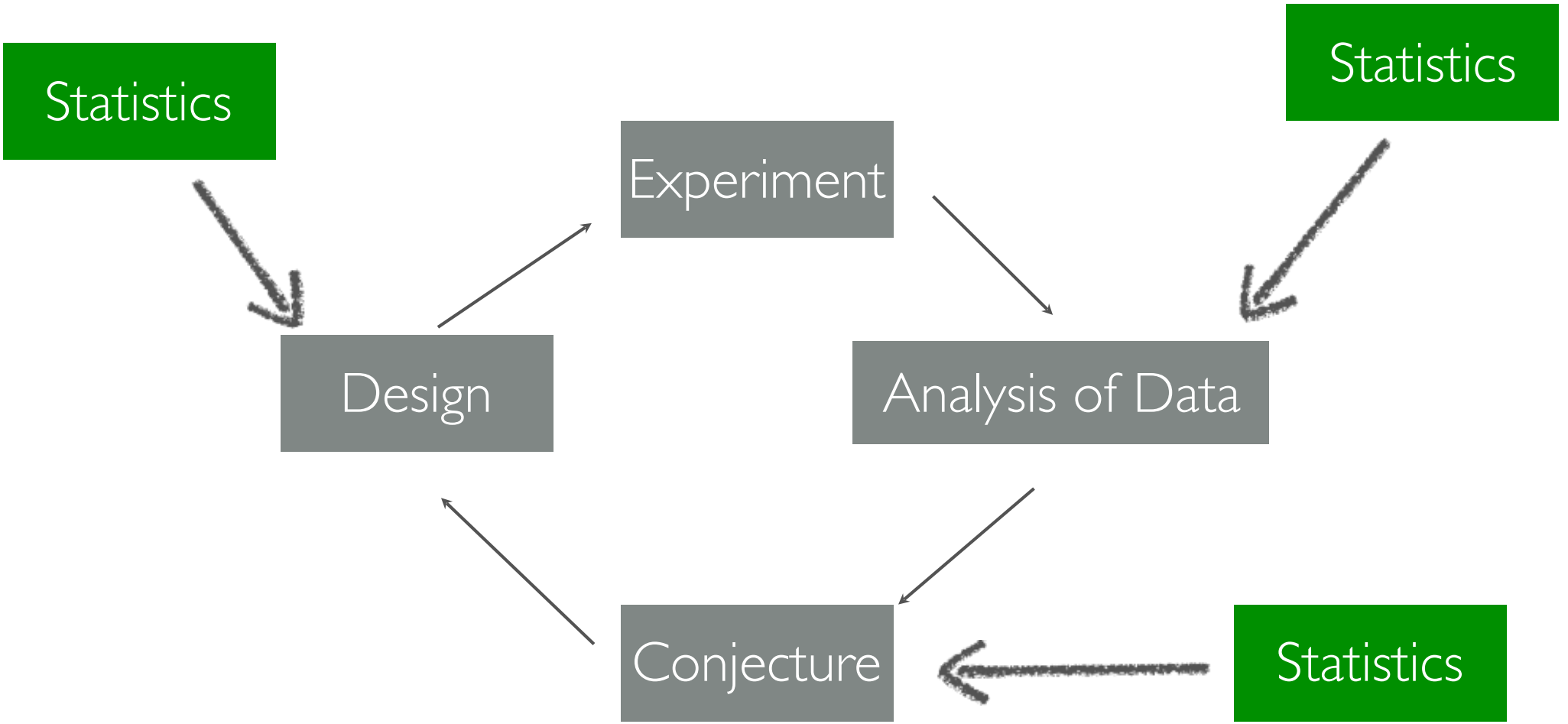
Response Surface Methodology

What Is Design of Experiments? Part 2

https://www.youtube.com/watch?v=hTviHGsl5ag

INTRODUCTION

Connected to Scientific Method



INTRODUCTION

- When you repeat an experiment you won't get the same response on two different occasions.
- Observation = true response + error
- The observations we get by repeating an experiment differ.

INTRODUCTION

- Good experimental design helps protect real effects from being obscured by experimental error.
- Designed experiments can increase signal-to-noise ratio.
- Statistical analysis provides measures of precision of estimated quantities under study.

Pathway to a Designed Experiment

OS or Other settings
Design of an Experiment

1 Experimental Design setting
2 Single, two or more than two factors
  Treatment design, environmental design…
3 Conduct the experiment, take observations …
4 Identify the influential factors alone/jointly
  Investigate for
  - Detection of factors' effects
  - Estimation of their effects

5 Estimation (continued)
  - Identify or estimate the optimal level/combination of trt. factors' levels
  - Inverse problem:
    For a desired response, estimate/identify the trt. factors' levels

6 Conclude OR go to Step 1

INTRODUCTION

- What is the optimal measurement strategy?
- Suppose that we want to measure mass of two apples A and B using an old-fashioned two-pan balance scale.
- Should the apples be weighed one at a time?

INTRODUCTION

- (Hotelling, 1944) Let σ^2 be the variance of individual weighings of two objects.

This apple has weight w_1

This apple has weight w_2

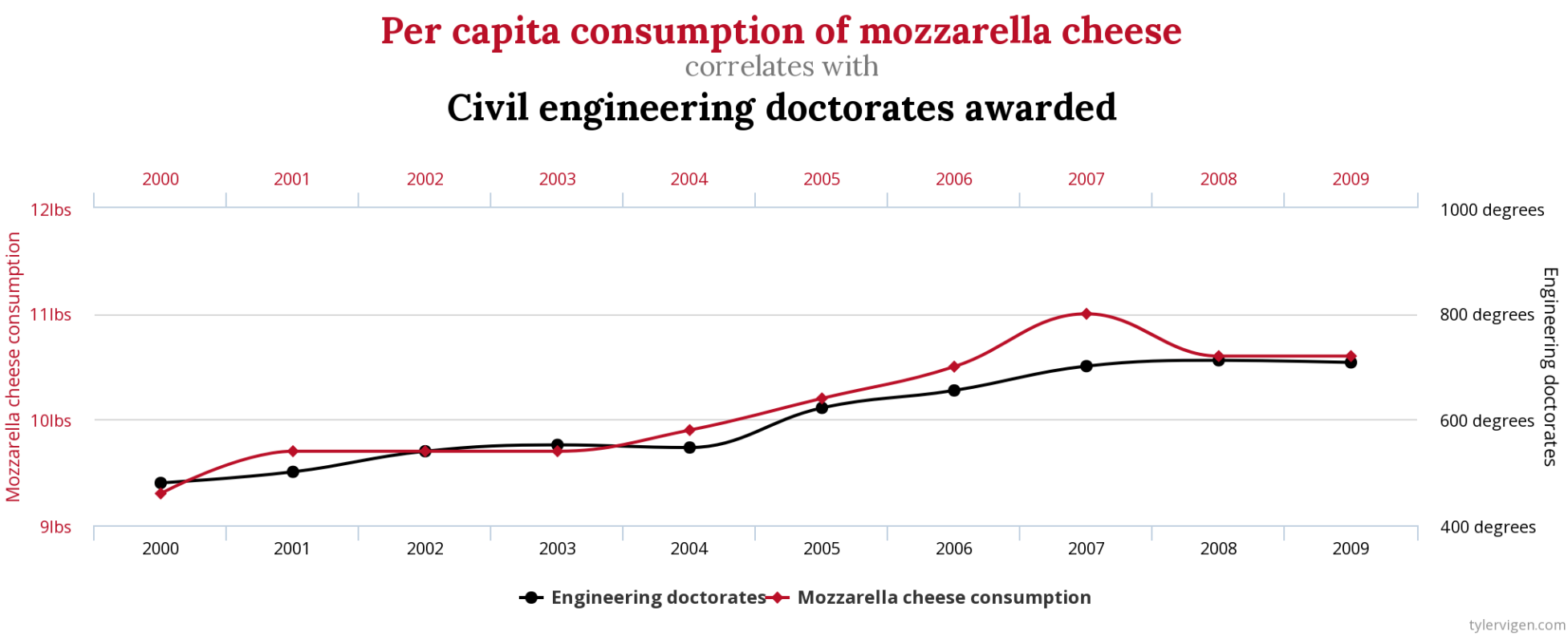
- Weigh two objects together in one pan to obtain the sum of the two weights.

- Weigh two objects in opposite pans to obtain the difference between the two weights.

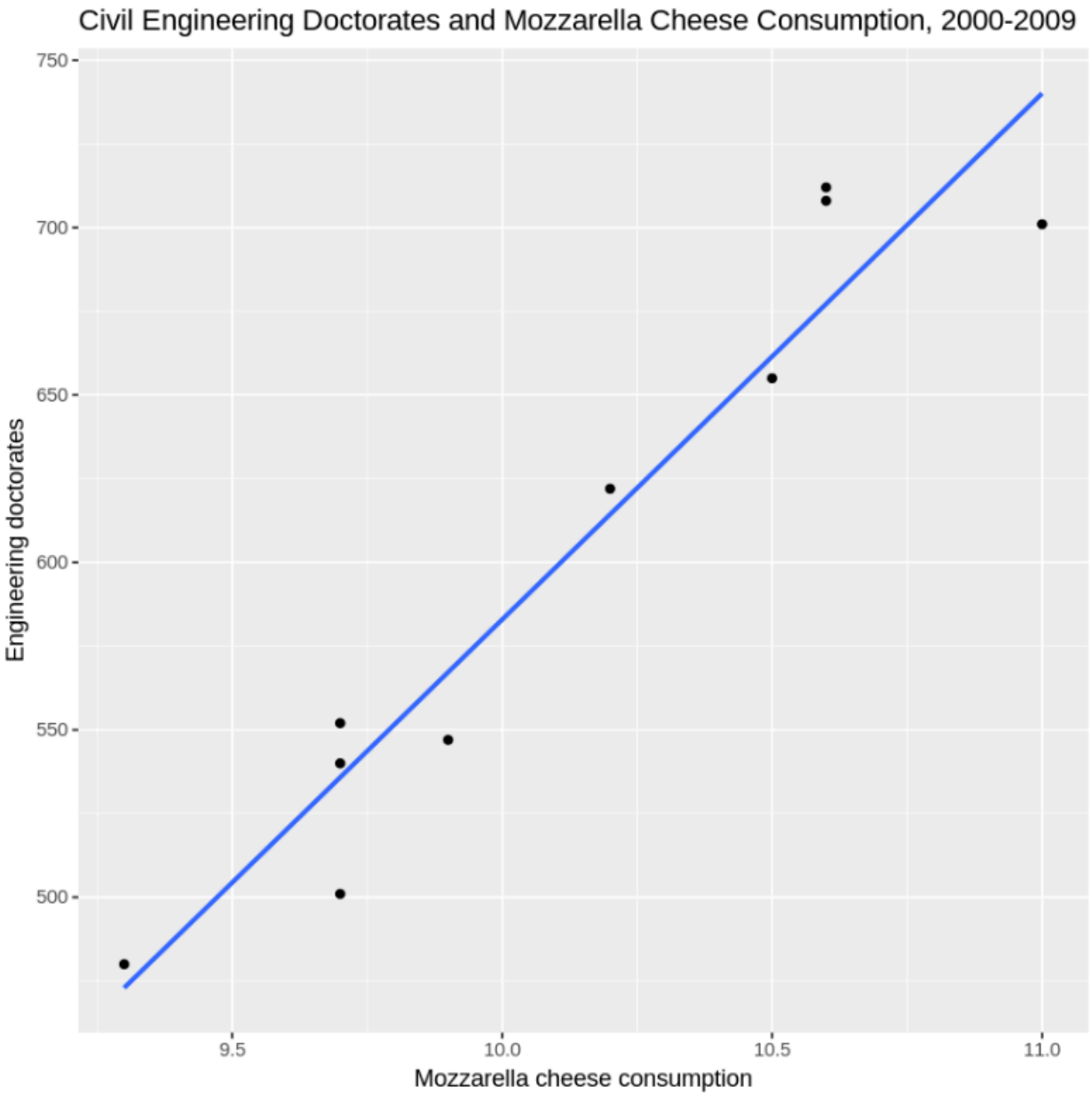
INTRODUCTION  

If the objects were weighed one at a time then how many weighings would be required to achieve the same precision (standard error) as the preceding design?

INTRODUCTION



INTRODUCTION



- A major issue in experimentation is confusion of correlation with causation.
- $R^2 = 0.919$ and p-value of slope is 1.22×10^{-5}.
- Does increase in cheese mozzarella cheese consumption increase the number of engineering doctorates awarded?

OVER

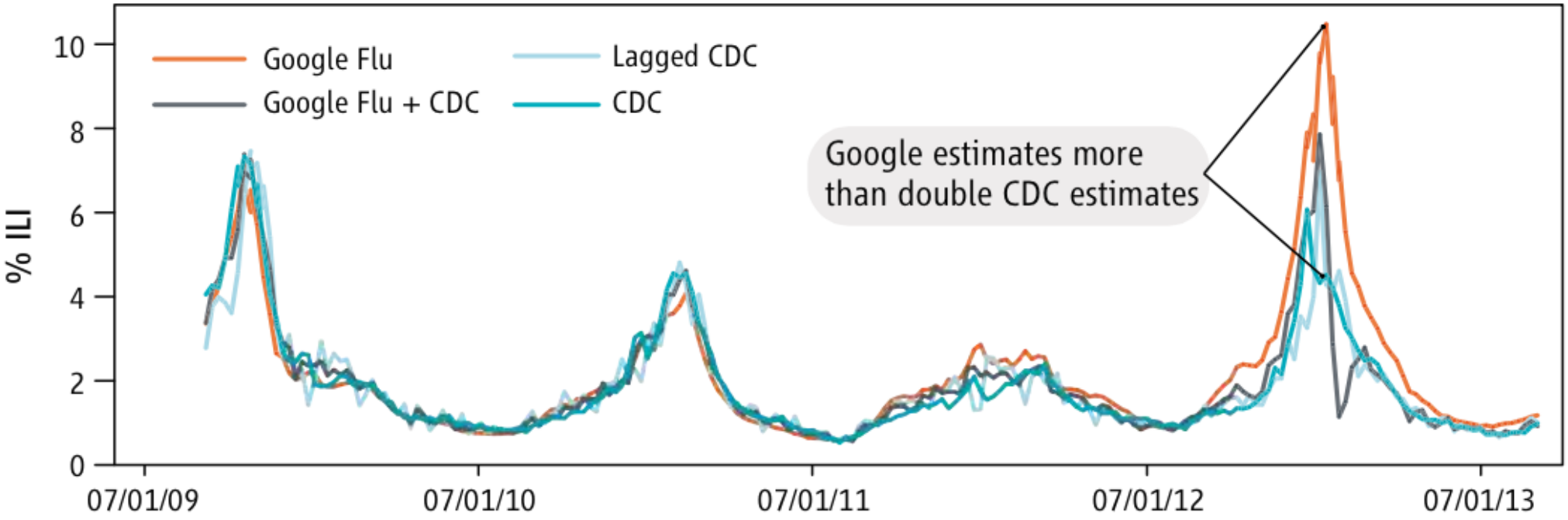
INTRODUCTION

- Data is usually very expensive.
- In a clinical trial the average per patient cost is between $5500-$7600.
- Statistics can help unfold what's going on in the lab or production facility.

INTRODUCTION

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

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



The data collection method has an impact on the quality of conclusions drawn from the data.

ABRAHAM WALD AND THE MISSING BULLET HOLES

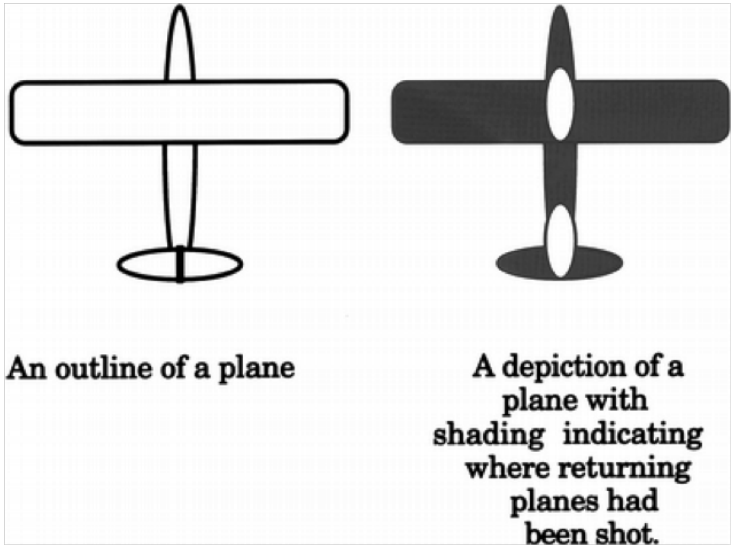
- During World War II he spent much of his time in the Statistical Research Group (SRG). A classified program that assembled the best American statisticians to the war effort.
- The SRG was in an apartment building in NYC a few blocks from Columbia U.

ABRAHAM WALD AND THE MISSING BULLET HOLES

- The SRG was a very influential group and the military frequently listened to their advice.
- Wald at the time was still an "enemy alien", he was not technically allowed to see the reports he was producing.

ABRAHAM WALD AND THE MISSING BULLET HOLES

The military supplied the SRG with some data



Design of Experiments and Observational Studies

Introduction

OS

U1, ..., Un

Observe for (X1, X2, ..., Xp, Y)

DS

U1, ..., Un

Observe for (X1, X2, ..., Xp)

Apply treatments (U s, randomization)

Measure/Observe (Y)

Inference domain

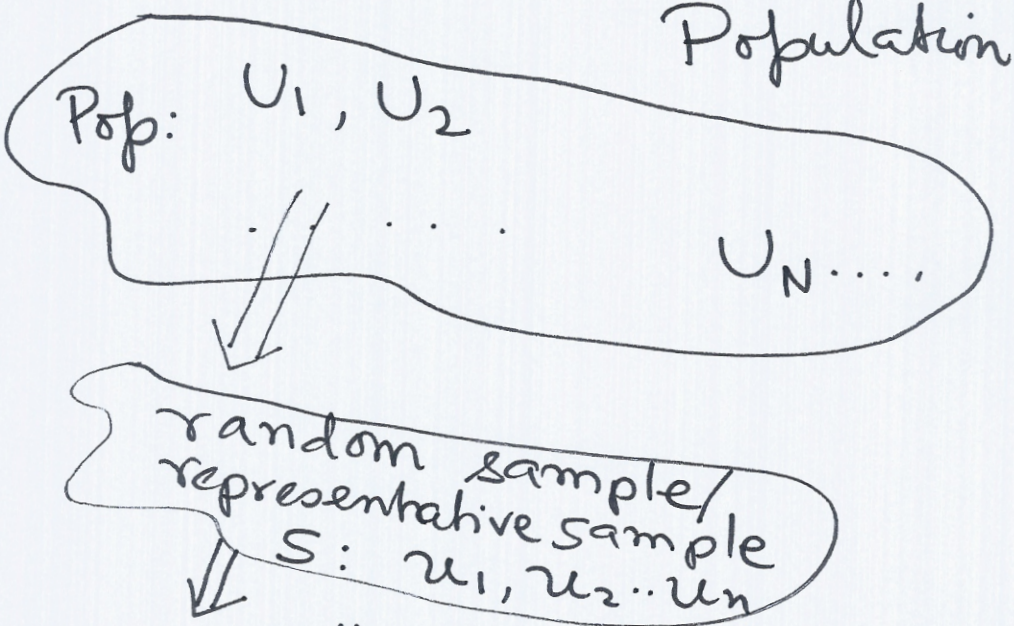
Association

Causation

Observational Study and Designed Study

↳ incl. survey

- Population
  - ↑ features of this pobn : Objective
  - ↳ Get data ← sample ← Pobn
    - ↓ Analyze
    - ↓ Infer



Collect / Get / Generate data

Observational study:

Researcher observes S: $u_1, u_2 .. u_n$ on $X_1 \dots X_p, Y_1 \dots$ as they are

Survey:

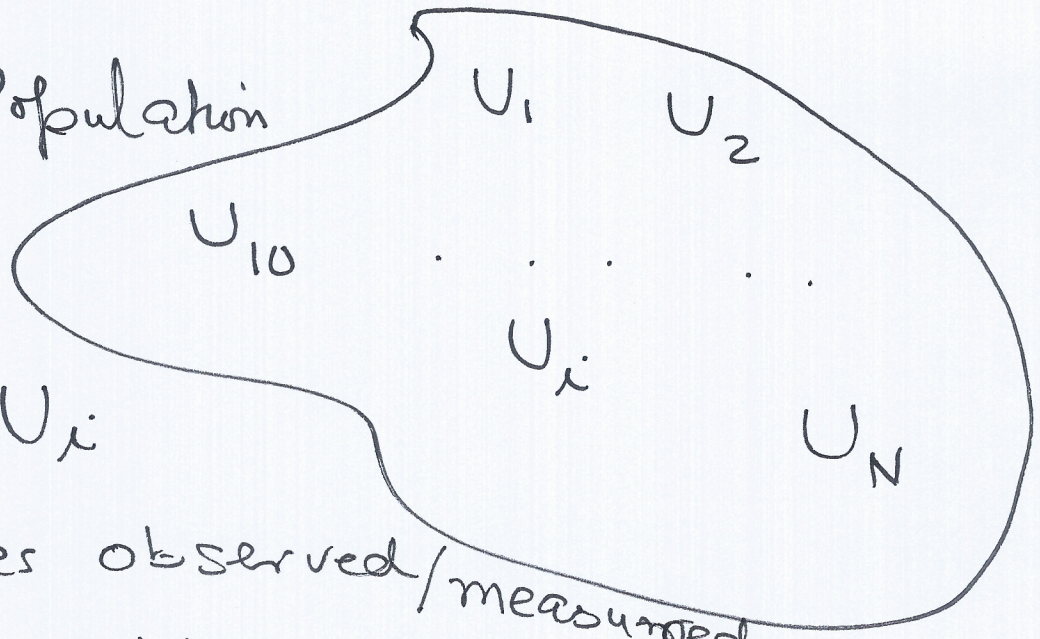
$\Rightarrow$ Response of $u_1, \dots u_n$ to interviews questions

- No factor → units assignment done

Designed study:

Researcher intervenes via a
- factor → units assignment

Let



Unit U_i

variables observed/measured

X_1	X_2	..	X_p	Y
X_{1i}	X_{2i}		X_{pi}	Y_i

$U_1, U_2 \ldots U_n$ randomly selected from P. $(i=1,2\ldots n)$

Situation 1 Observe P as is

Using a random sample $\{U_1, U_2 .. U_n\}$ and infer on P for the association between $X_1, X_2, \ldots Y \ldots$

- no prior knowledge of the assignment between the unit and a level of X_j $j=1,2,\ldots p$

Situation 1 is observational study.

Approach 1 : Examine the associations between $X_1, \ldots X_p, Y$..

Let $P = \{$ children, age 10–20 yrs, in a given district $\}$

H : Height (cm)
W : Weight (kg)

(i) $H_{10} \longleftrightarrow W_{10}$ (associated)

age $\longrightarrow H_{10}$ cause, effect ?.

(ii) H_{10} W_{10} (associated)

$Z_1, Z_2 \ldots$ latent varis

(iii) $\{H_{10}, W_{10}\} \longrightarrow W_{20}$

Can these be justified as causes? effect

predictor — ?.

(iv) X : smoking behavior $\{S, NS\}$
Y : BP measurements
in a specified popn.

Data : $\{ X_i, Y_i, i = 1 \ldots n$ from $P \}$
(U_i random)

Is X a cause for Y ?

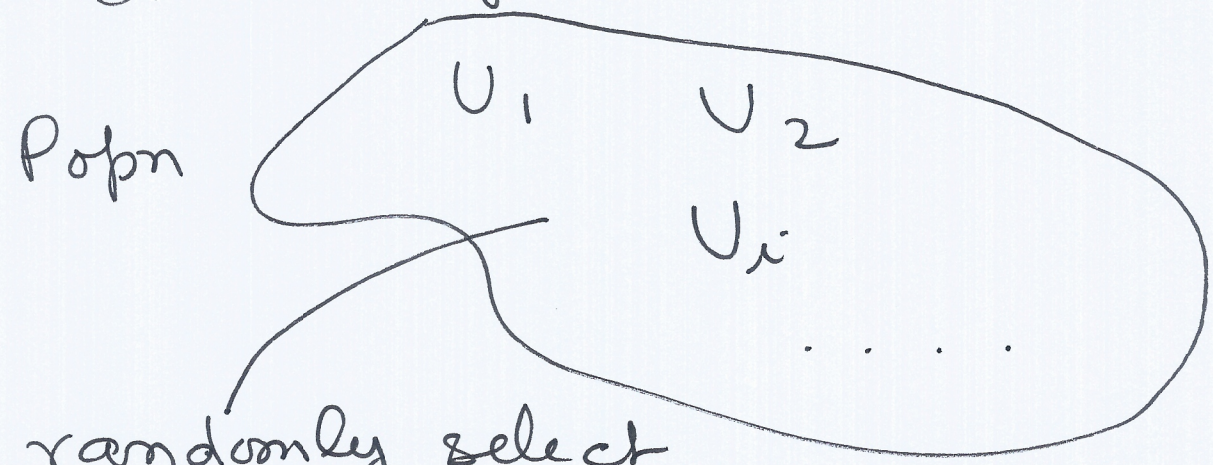
- Case of observational study as X_i was found on U_i, not assigned to U_i.
- Inferences on association
  - not on causation.
- Association is expressed as between
  - correlation — quantitative vars
  - contingency χ^2 — qualitative vars
  - variance ratio — a qualitative & a quantitative variable

Approach 2 — Where justification is made for $X_1 \ldots X_p$ as predictors for Y as effect / response

Regression relationship is obtained.

Situation 2 Designed study/ Experiment

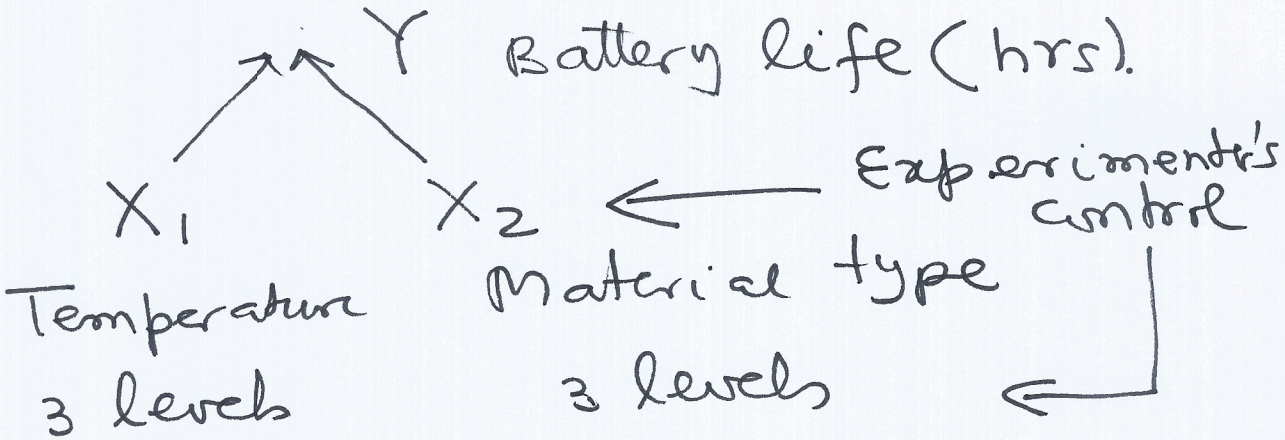
- Designed to intervene to change the response by controlling $X_1, \ldots X_p$ (factor or variable).



- randomly select $u_1, u_2 \ldots u_n$
- Apply chosen levels of $X_1 \ldots X_p$ to u's
  → creates responses y's.

i.e. $X_1 \ldots X_p$ are outside the popn. (unlike the observational study case).

- Example: Battery Design (Chapter 3) DOX Montgomery.



X_1, X_2 : causes

Y : Effect

Association between X_1 & X_2 is meaningless as their combinations are chosen by the experimenter.

Here one models Y interms of X_1 & X_2:

$$Y_i = f(X_{1i}, X_{2i}) + \epsilon_i$$

$i = 1, \ldots, n$

$f(\,)$ is explored empirically and modified/updated.