Introduction to Causal Inference

Correlation ≠ Causation

A Pacific island tribe observed that healthy people had lice, while sick people did not. Accordingly, they believed that lice infestations are good for one's health.

Of course, lice do not cause good health

Causality runs the other way: lice abandon a fevered body to seek a healthier host

Tim Maudlin (Boston Review, Sept 4 2019)

Correlation → Causation?

CNN @CNN · Nov 5

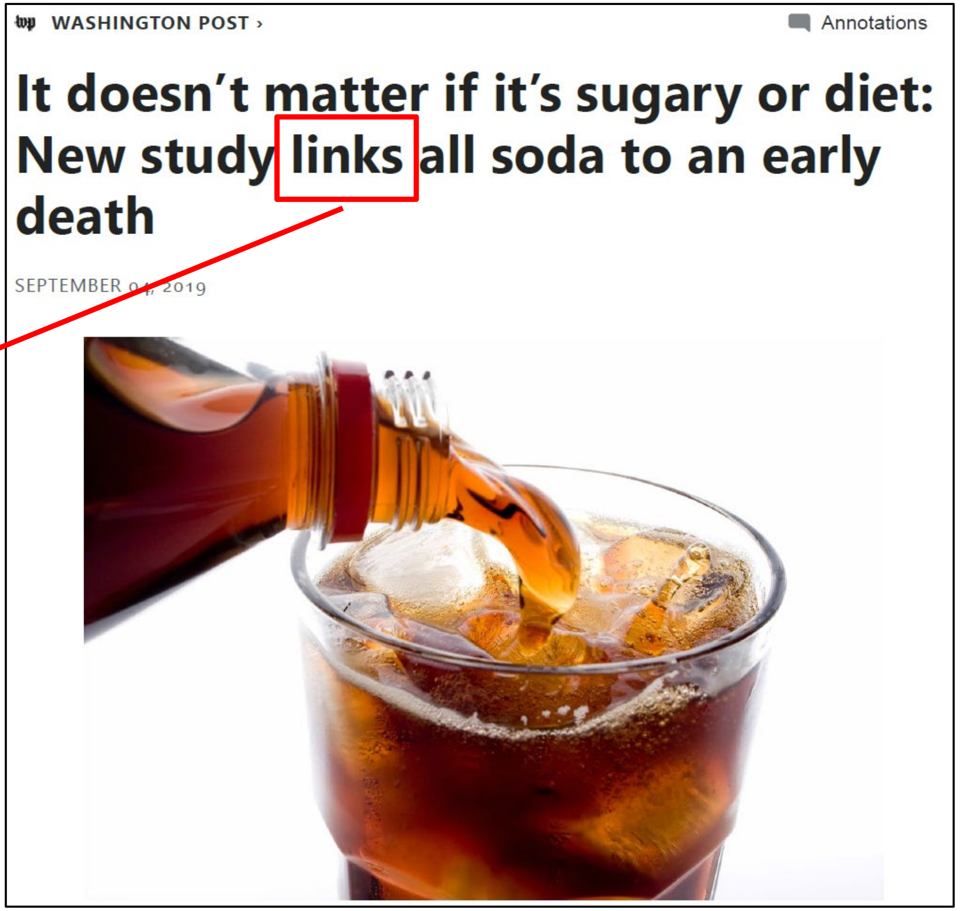
Wondering whether it's worth going for a little jog? Get those sneakers on -- a new study shows that any amount of running lowers the risk of premature death.

Even a little running cuts risk of premature death, new study shows

Wondering whether it's worth going for a little jog? Get those sneakers on -- a new study shows that any amount of running lowers the risk of ...

edition.cnn.com

Correlation → Causation?


WASHINGTON POST ›

Annotations

It doesn’t matter if it’s sugary or diet: New study links all soda to an early death

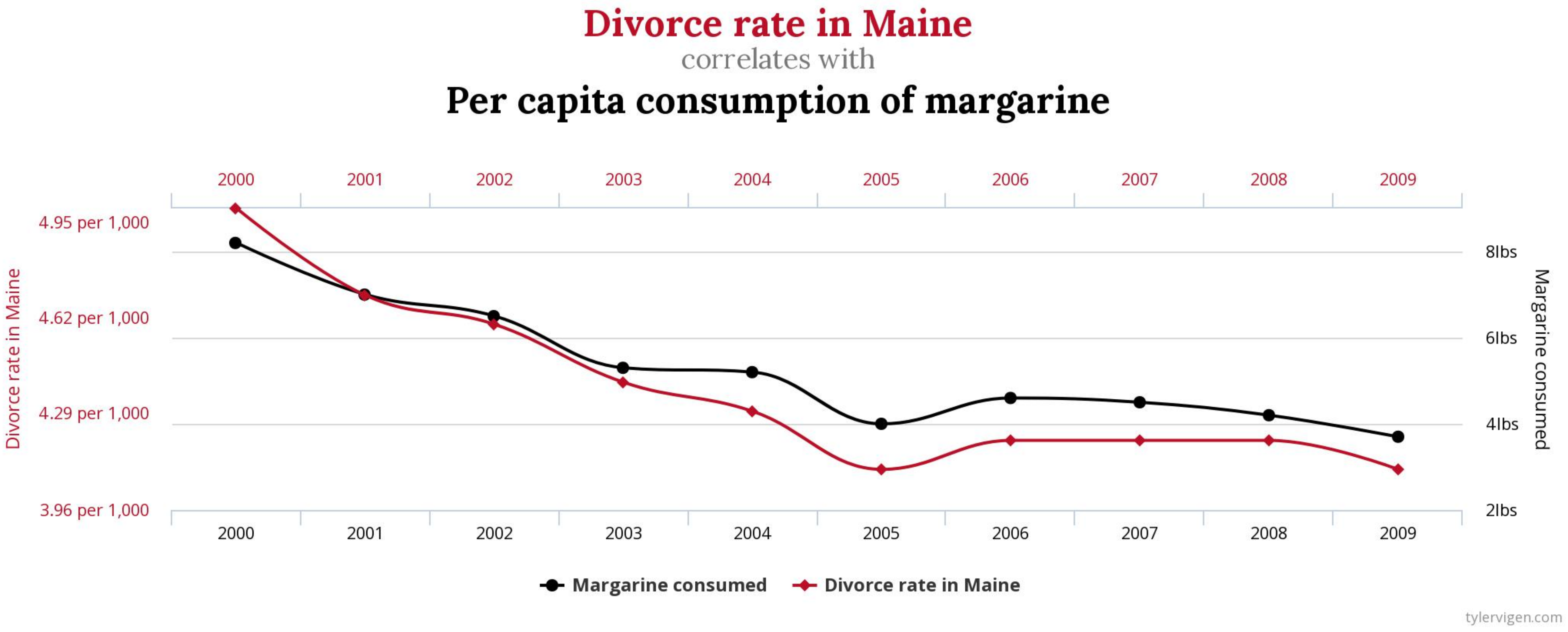
SEPTEMBER 04, 2019

People tend to **interpret correlations as causation**

A Simple Causal Graph

$$D \xrightarrow{?} Y$$

Spurious Correlations



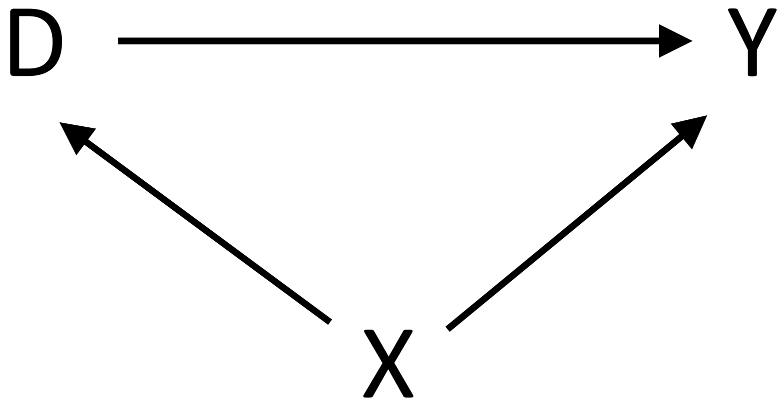
Reverse Causality

$$D \longleftarrow Y$$

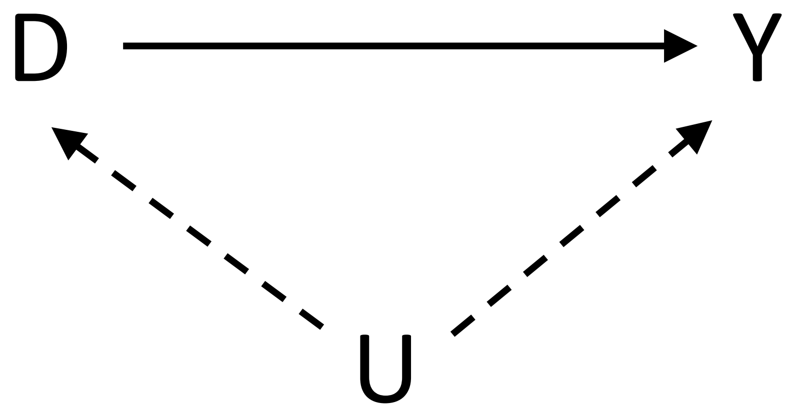
Simultaneity

$$D \longleftrightarrow Y$$

Omitted Variable Bias (or Selection Bias)



Omitted Variable Bias (or Selection Bias)



Key challenge of causal inference is to solve the problem of unobserved confounders

Should You Get Health Insurance (or Not)?

Suppose the decision of getting health insurance is a matter of …

- Insurance premium
- Quality of care

You observe a strong, long-term correlation between having health insurance and positive health outcomes

Should you get health insurance?

Many Possible Mechanisms, Which One Is It?

Being uninsured may lead to poorer health outcomes → causal effect

Poor health may lead to loss of health insurance → reverse causality

Uninsured may be more adventurous or less concerned about their health → selection bias

All of these hypotheses could account for the observed correlation

Why Causality Matters

Identifying *true cause* is crucial because each explanation gives different advice

Machine learning and large data sets may help us *predict* what will happen

But even the most sophisticated machine learning tools cannot make the data tell us *why*

> "We live in an era that presumes Big Data to be the solution to all our problems"
>
> (Pearl, The Book of Why)

How Can We Learn About Causality?

What is the causal impact of health insurance on *your* health?

Causal effect = (your health outcome *with* insurance) − (your health outcome *without* insurance)

... holding **all other factors constant**

- Ideally, we would clone you and measure your and your clone's health under different insurance conditions

- In reality, we either observe you with or without health insurance, but never both

What is Causal Inference?

Counterfactual outcome = What would have happened had this one aspect not changed

We never observe the counterfactual → **fundamental problem of causal inference** (Holland, 1986)

Causal inference is about how we can use **data + assumptions** to learn something about the counterfactual

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information

Selection Bias

Does Health Insurance Make People Healthier?

- Universal health care provides quality medical services to all citizens
- The Affordable Care Act (“Obamacare”) requires U.S. citizens to have health insurance (or pay a tax penalty)
- But many U.S. citizens are still uninsured and have to rely on emergency departments

What is the causal effect of health insurance on people’s health?

Simple Difference in Mean Outcomes

We cannot observe a person both with and without health insurance at the same time

- Thus, we cannot calculate the causal effect of health insurance *on that person's* health

What if we compare the difference in average health outcomes between the insured and uninsured?

- Is this a *ceteris paribus* comparison?

Simple Difference in Mean Outcomes

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	$\boxed{.31}$ (.03)	4.02 [.92]	3.62 [1.01]	$\boxed{.39}$ (.04)
Sample size	8,114	1,281		8,264	1,131	

Those with health insurance are healthier than those without

Ceteris paribus?

Selection Bias

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Insured are better educated

Insured are more likely to have a job

Selection Bias

- Variables like education and employment are correlated with health insurance and health status

- Large differences in observed variables between the two groups are an indication that there are also **relevant differences in unobserved variables**

What are we measuring when we compare the average health status of the insured and uninsured?

Notation

Population of interest: All individuals i who are eligible for health insurance

Observed outcome: Y_i is the health status of individual i (on a scale of 1 to 5)

(Binary) treatment:

$$D_i = \begin{cases} 1 & \text{if } i \text{ is insured} & \text{(treated)} \\ 0 & \text{if } i \text{ is uninsured} & \text{(control)} \end{cases}$$

Pretreatment covariates: X_i (gender, education, income etc.)

Potential Outcomes Model

Each *individual* has **two potential outcomes:**

$$\text{Potential outome} = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

- Y_i^1 is the health of individual i if she is insured
- Y_i^0 is the health of the *same individual* i if she is not insured

For any individual, we can only **observe one potential outcome**:

- Each individual either has or doesn't have health insurance
- The *counterfactual* outcome is not observed but we can imagine it

Potential Outcomes Model

For each individual, the **causal effect** of treatment D is the difference between two potential outcomes:

$$\tau_i \equiv Y_i^1 - Y_i^0$$

Potential outcomes are ***hypothetical*** variables: (Y_i^1, Y_i^0)

Observable outcomes are ***factual*** variables: (Y_i)

The link between an individual's actual outcome Y_i and her potential outcomes is:

$$\underbrace{Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0}_{\text{"Switching equation"}} = Y_i^0 + \underbrace{(Y_i^1 - Y_i^0)}_{= \tau_i} D_i$$

Stable Unit Treatment Value Assumption (SUTVA)

No interference between individuals:

- Treatment assignment of one person does not affect that outcome of another person
- Possible violations: spillover effects between individuals

No variation in the treatment:

- Only one version of the treatment (treatment must be well defined)
- Possible violations: people receive different ‘doses’ of the treatment

Example – Student Health Insurance

Khuzdar is a visiting UM student who recently arrived from Kazakhstan

- Khuzdar has to decide whether to opt in to the university’s health insurance plan
- Khuzdar decides to take up UM’s health insurance

Maria is also coming to UM this year from Chile

- Maria must also decide whether to opt in to the university’s health insurance plan
- Maria decides not to get health insurance

Example – Student Health Insurance

Subject	Treatment	Actual outcome	Potential outcomes		Individual treatment effect
	D_i	Y_i	Y_i^1	Y_i^0	$\tau_i = Y_i^1 - Y_i^0$
Khuzdar	$\boxed{1}$	$\boxed{4}$	$\boxed{4}$	3	$\boxed{1}$
Maria	$\boxed{0}$	$\boxed{5}$	5	$\boxed{5}$	$\boxed{0}$

Outcome: Health status
(1 = ‘poor’ to 5 = ‘excellent’)

Causal Inference as a Missing Data Problem

Subject	Treatment	Actual outcome	Potential outcomes		Individual treatment effect
	D_i	Y_i	Y_i^1	Y_i^0	$\tau_i = Y_i^1 - Y_i^0$
Khuzdar	1	4	4	?	?
Maria	0	5	?	5	?

Outcome: Health status
(1 = ‘poor’ to 5 = ‘excellent’)

Simple Difference in Individual Outcomes

Khuzdar and Maria made different insurance choices

- The difference in actual outcomes between Khuzdar and Maria:

$$(Y_K|D_K = 1) - (Y_M|D_M = 0) = 4 - 5 = -1$$

- This suggests that health insurance causes health to become worse

What's wrong with this picture?

Selection Bias

Let's link actual and potential outcomes:

$$\underbrace{(Y_K|D_K = 1) - (Y_M|D_M = 0)}_{\text{Actual outcomes}} \quad = \quad \underbrace{Y_K^1 - Y_M^0}_{\text{Potential outcomes}}$$

Adding in $\underbrace{-Y_K^0 + Y_K^0}_{=0}$, we get:

$$(Y_K|D_K = 1) - (Y_M|D_M = 0) \quad = \quad \underbrace{Y_K^1 - Y_K^0}_{\substack{=1 \\ \text{Causal effect} \\ \text{for Khuzdar}}} \quad + \quad \underbrace{Y_K^0 - Y_M^0}_{\substack{=-2 \\ \text{Selection bias}}}$$

Average Treatment Effect

The **average treatment effect** (ATE) is the sum of the individual treatment effects divided by the number of people:

$$\text{ATE} \equiv \frac{1}{N}\sum_{i=1}^{N} \tau_i$$

This is equivalent to:

$$\frac{1}{N}\sum_{i=1}^{N} \tau_i = \frac{1}{N}\sum_{i=1}^{N} [Y_i^1 - Y_i^0] = \frac{1}{N}\sum_{i=1}^{N} Y_i^1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0$$

ATE is the difference in *population* means of the potential outcomes

Simple Difference in Mean Outcomes

Difference in group means

$$= Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0] \qquad \text{(avg. of actual outcomes)}$$

$$= Avg_n\left[Y_i^1|D_i = 1\right] - Avg_n\left[Y_i^0|D_i = 0\right] \qquad \text{(avg. of potential outcomes)}$$

$$\neq Avg_n\left[Y_i^1 - Y_i^0\right]$$

For simplicity, assume *constant* (κ) treatment effects:

$$Y_i^1 = Y_i^0 + \kappa \quad \text{for all } i$$

Simple Difference in Mean Outcomes

Difference in group means

$$= Avg_n\left[Y_i^1 | D_i = 1\right] - Avg_n\left[Y_i^0 | D_i = 0\right]$$

$$= \left\{\kappa + Avg_n\left[Y_i^0 | D_i = 1\right]\right\} - Avg_n\left[Y_i^0 | D_i = 0\right]$$

$$= \underset{\substack{\downarrow \\ \text{Average treatment effect}}}{\kappa} + \underbrace{\left\{Avg_n\left[Y_i^0 | D_i = 1\right]\right\} - Avg_n\left[Y_i^0 | D_i = 0\right]}_{\text{Selection bias}}$$

Solving the Selection Problem

Selection bias is the inherent differences between the two groups if both were uninsured

- *Observed* differences (e.g. education)
- *Unobserved* differences (e.g. exercise habits)

Selection problem is (relatively) easy to fix if the only source of selection bias are observed differences between groups

- But if there are large observed differences, we should be wary of unobserved differences

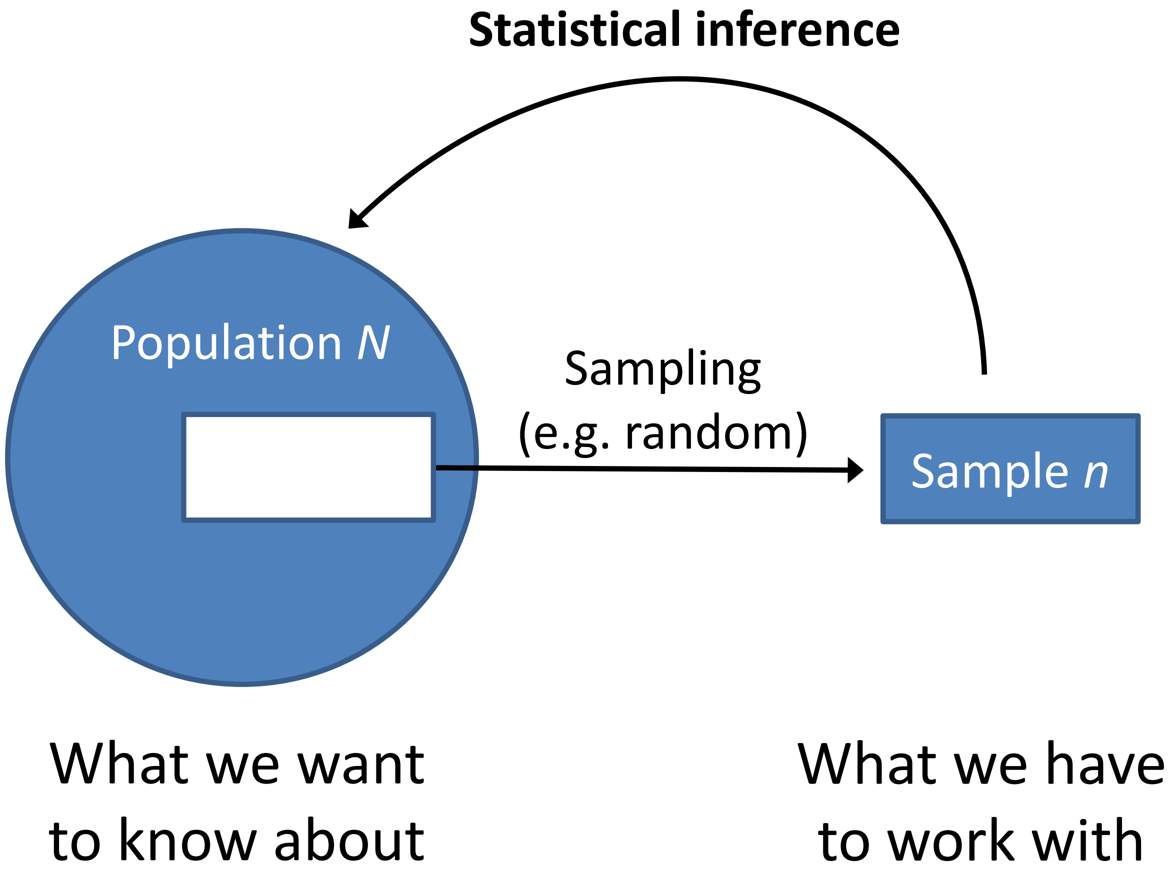
Main challenge of causal inference is to eliminate selection bias from unobserved differences

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information

Statistical Inference

Statistical Inference



Statistical Inference

1. Estimate a population value

- What is the best estimate of the causal effect of health insurance on health?

2. Estimate precision of an estimate (standard error)

- What range of values can we be confident contains the true value for the causal effect of health insurance on health?

3. Hypothesis testing (p-value)

- Could the insurance effect we observed be attributed to chance?

Average Causal Effects

Average treatment effect (ATE):

$$Avg_N[\tau_i] = \frac{1}{N}\sum_{i=1}^{N}[Y_i^1 - Y_i^0]$$

- ATE is the average change in outcomes if *everybody* in the population were to go from untreated to treated
- ATE is the average of the unit-level treatment effects
- ATE is not a quantity we can calculate because we would need the counterfactual outcome for each individual—but we can *estimate* it

Average Causal Effects

Conditional average treatment effect (CATE) for a subpopulation:

$$Avg_{N_x}[\tau_i | X_i = x] = \frac{1}{N_x} \sum_{i:X_i=x} [Y_i^1 - Y_i^0]$$

where N_x is the number of individuals in the subpopulation (e.g. women)

Average treatment effect on the treated (ATT):

$$Avg_{N_t}[\tau_i | D_i = 1] = \frac{1}{N_t} \sum_{i:D_i=1} [Y_i^1 - Y_i^0]$$

where $N_t = \sum_i D_i$

Random Sampling

In estimating ATE, we rarely observe the entire population of interest → sample

We can draw a *random sample* of individuals and calculate the average outcome among those sampled

The sample average is a quantity that varies from sample to sample

But the *expected value* of the sample average equals the average of the population from which the sample is drawn

Expectations

$E[Y_i]$ is the expected value of Y_i when one unit is sampled at random

$E[Y_i]$ is also called the *population mean* because it's the weighted average of all possible values that Y_i can take

- e.g., the expected outcome of a die roll is 3.5 (1/6*1+1/6*2+ etc. = 3.5)

Note that $E[Y_i]$ does not reference the sample size because it's a population quantity $\rightarrow$ parameter

Unbiasedness of the mean of a random sample: $E[\bar{Y}] = E[Y_i]$

- where $\bar{Y}$ is short for $Avg_n[Y_i]$

Conditional Expectations

$E[Y_i|X_i = x]$ is the expected value of Y_i when one unit is sampled at random among those whose value of $X_i = x$

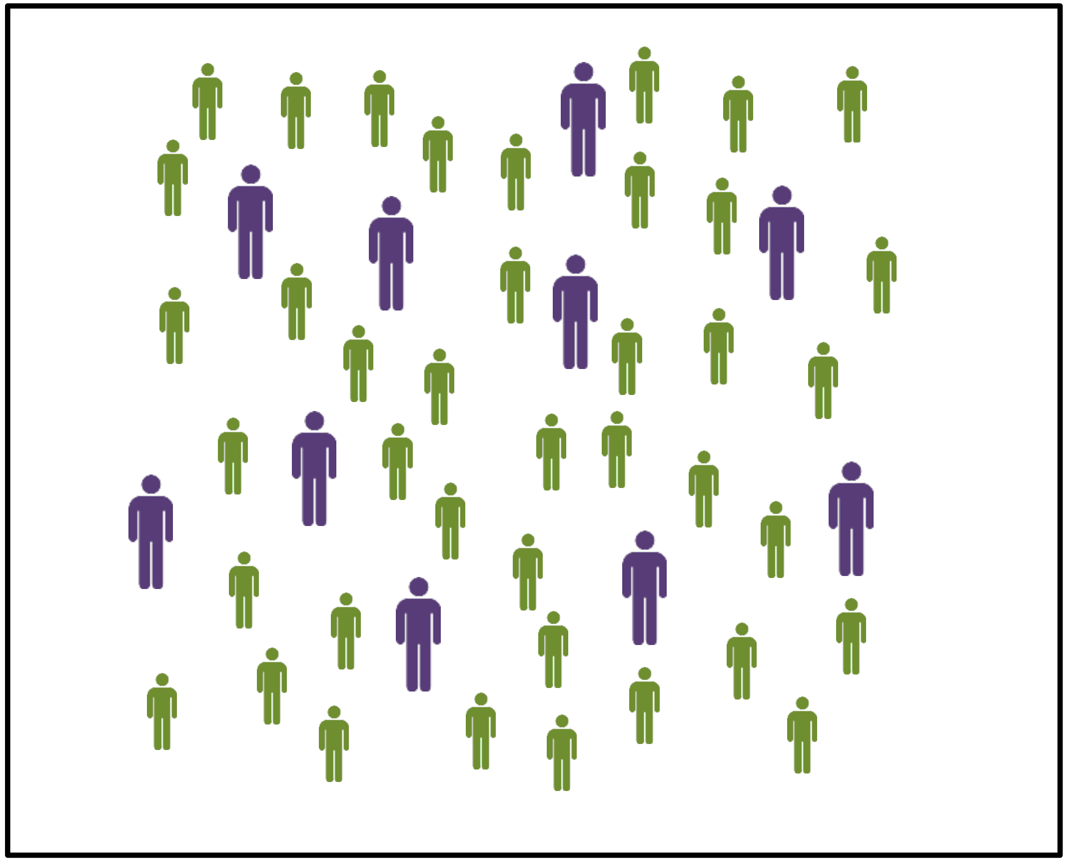
- Conditional expectations refer to subgroup averages
- Conditional expectations are easy to understand when working with variables that are observable

More mind-bending are expressions like $E\left[Y_i^1|D_i = 0\right]$

- Expectation of the potential outcome Y_i^1 when one unit is selected at random among those who are *not* treated

The Law of Large Numbers

Population of eligible individuals:



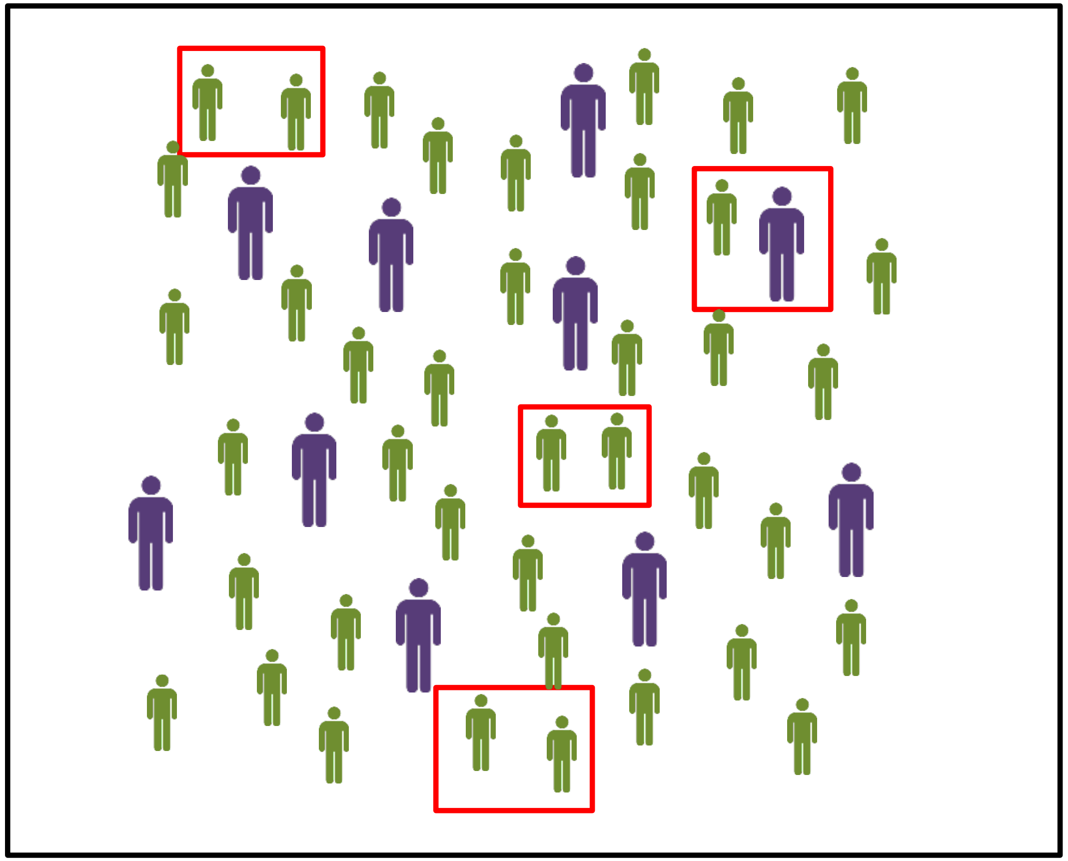
25% PURPLE individuals

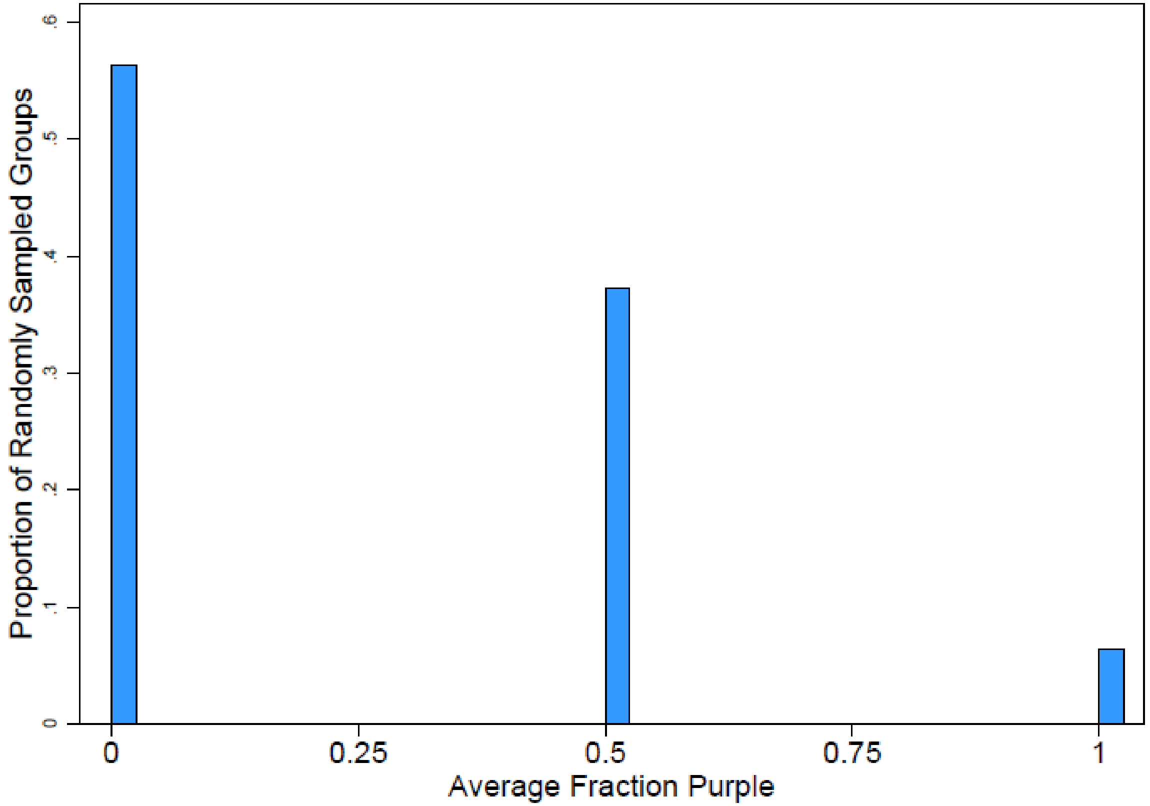
If you chose one individual at random, probability that she is PURPLE: 0.25

However, any one individual (chosen at random) is either PURPLE or GREEN

The Law of Large Numbers

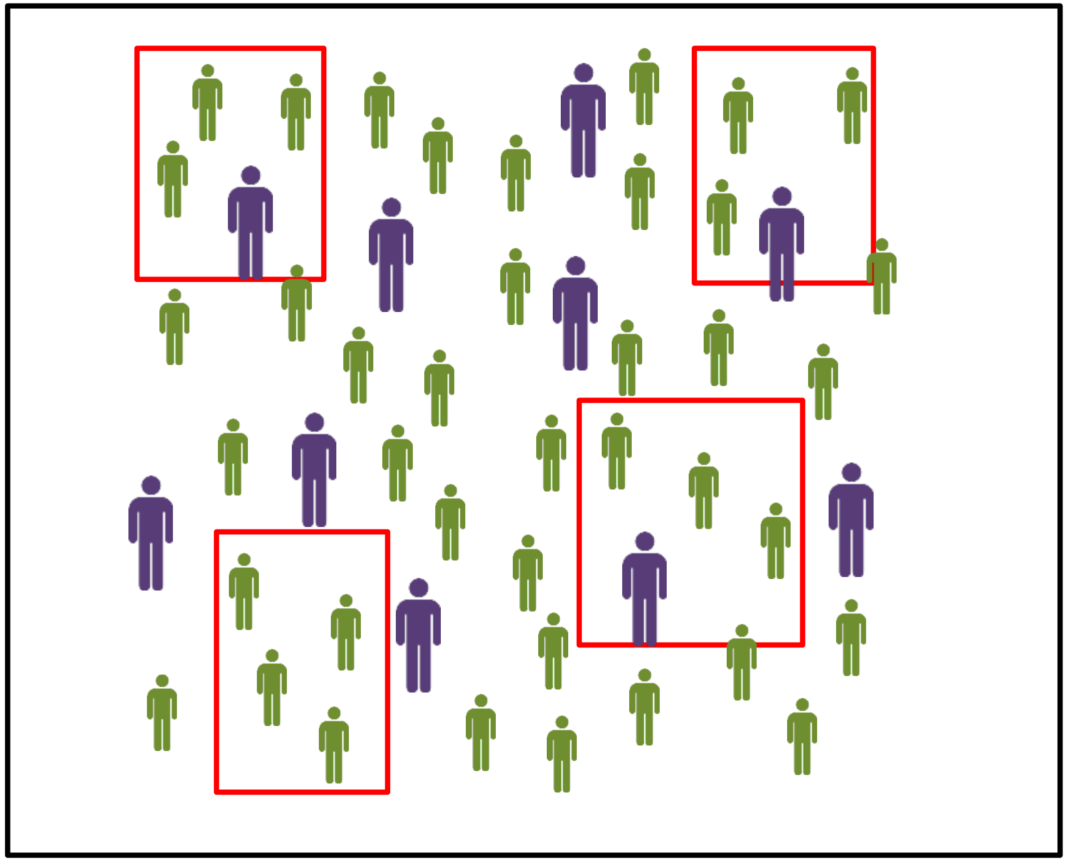
When you randomly sample groups of 2:

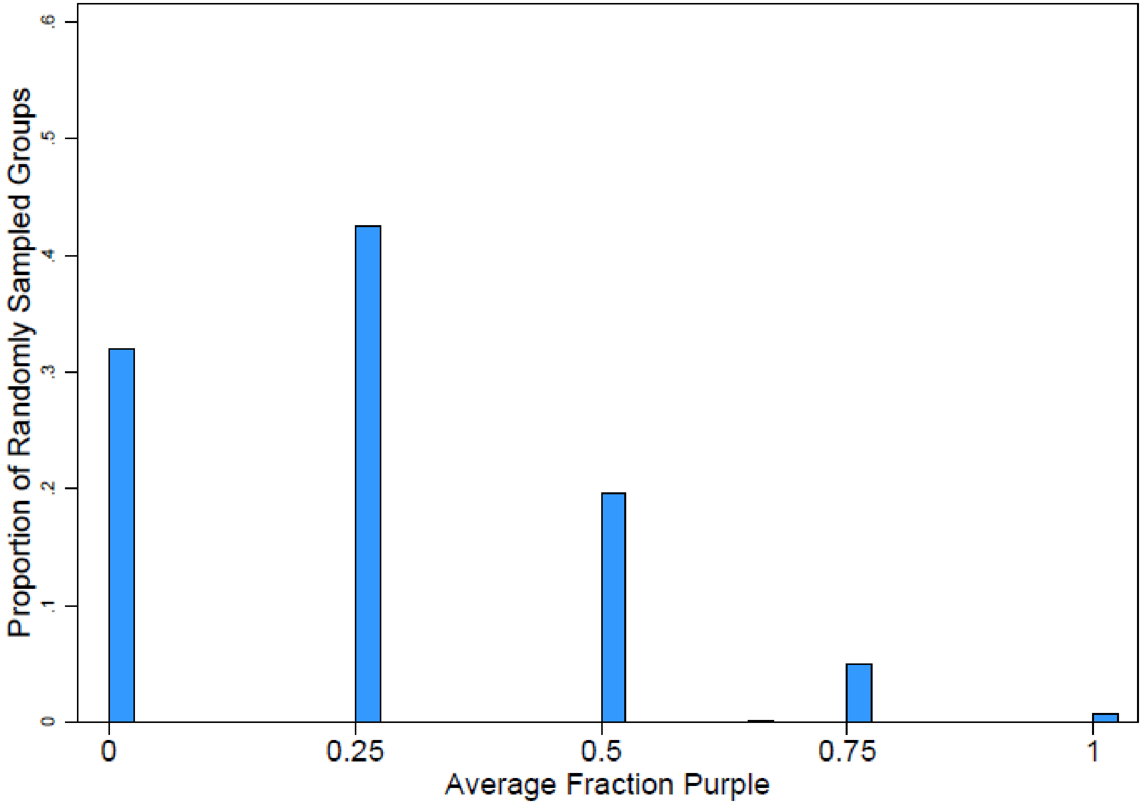




The Law of Large Numbers

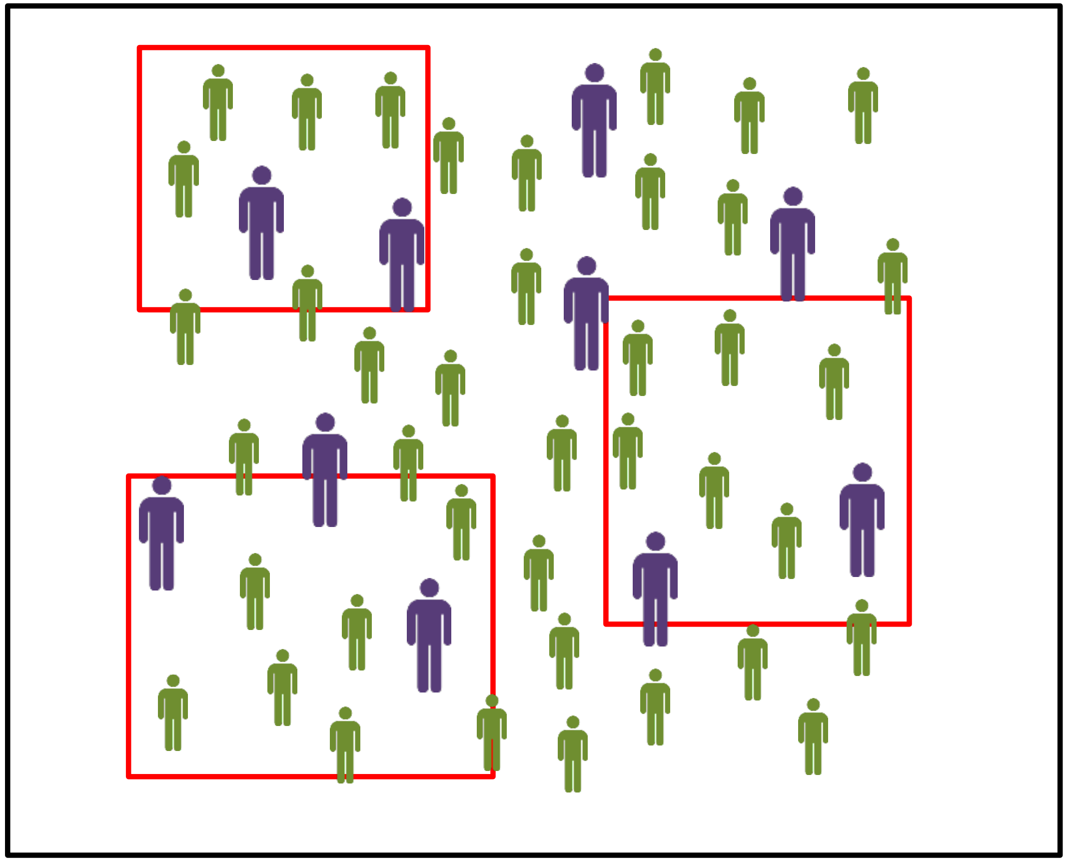
When you randomly sample groups of 4:

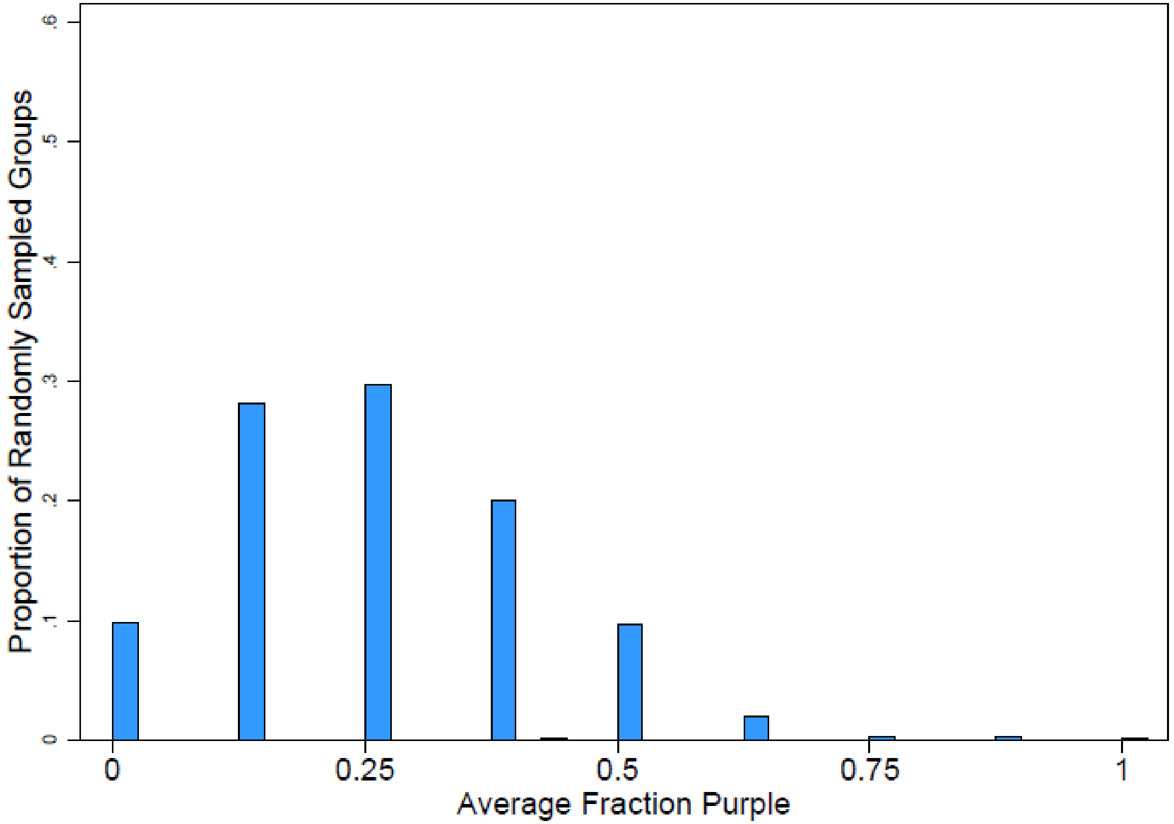




The Law of Large Numbers

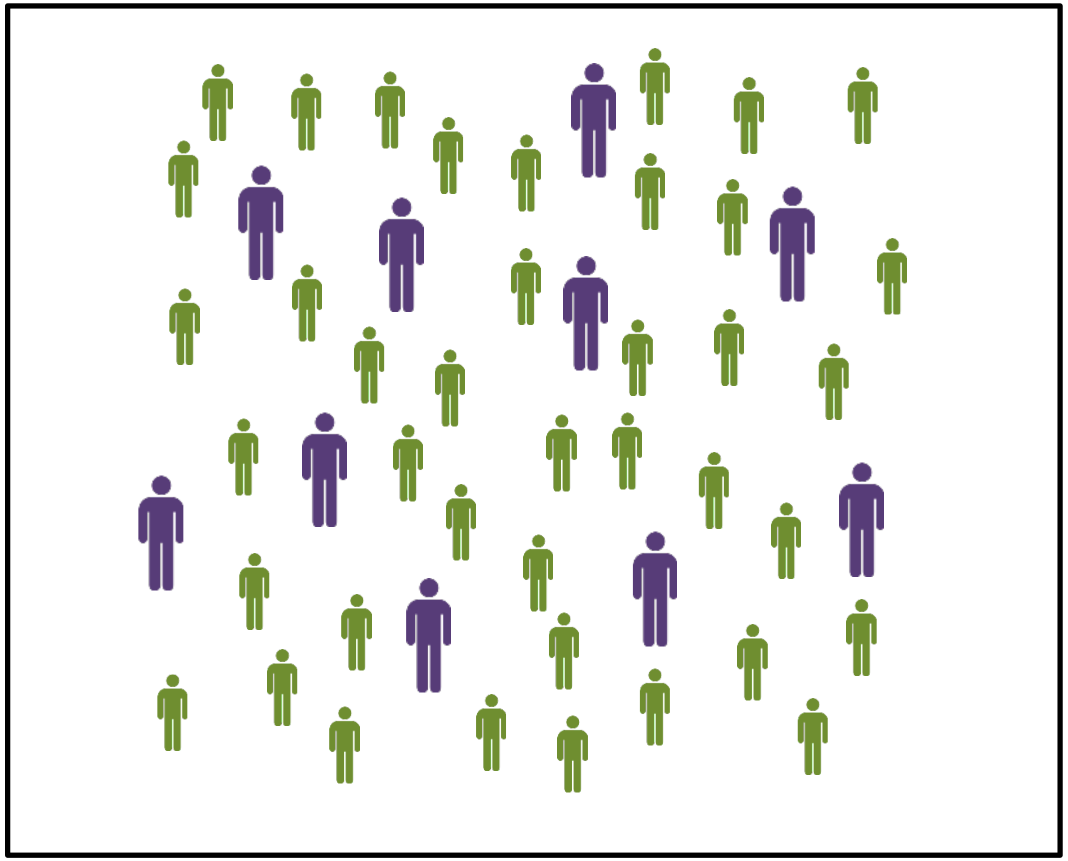
When you randomly sample groups of 8:

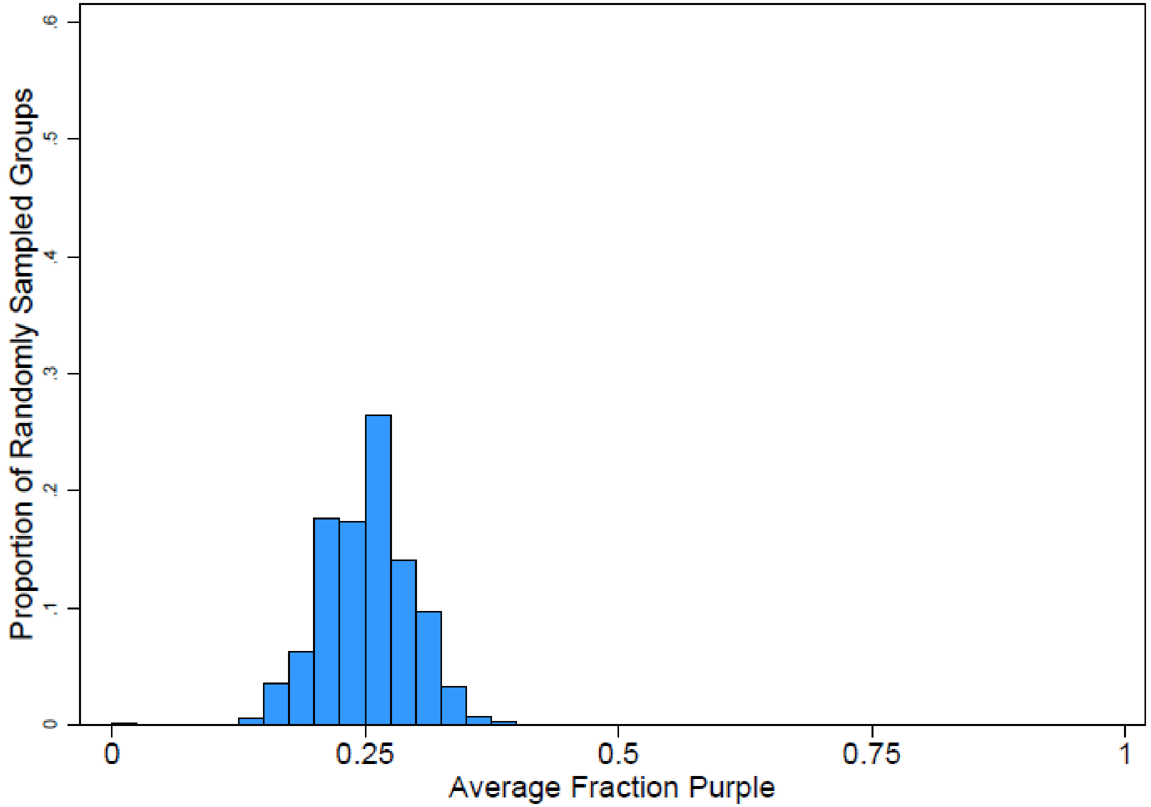




The Law of Large Numbers

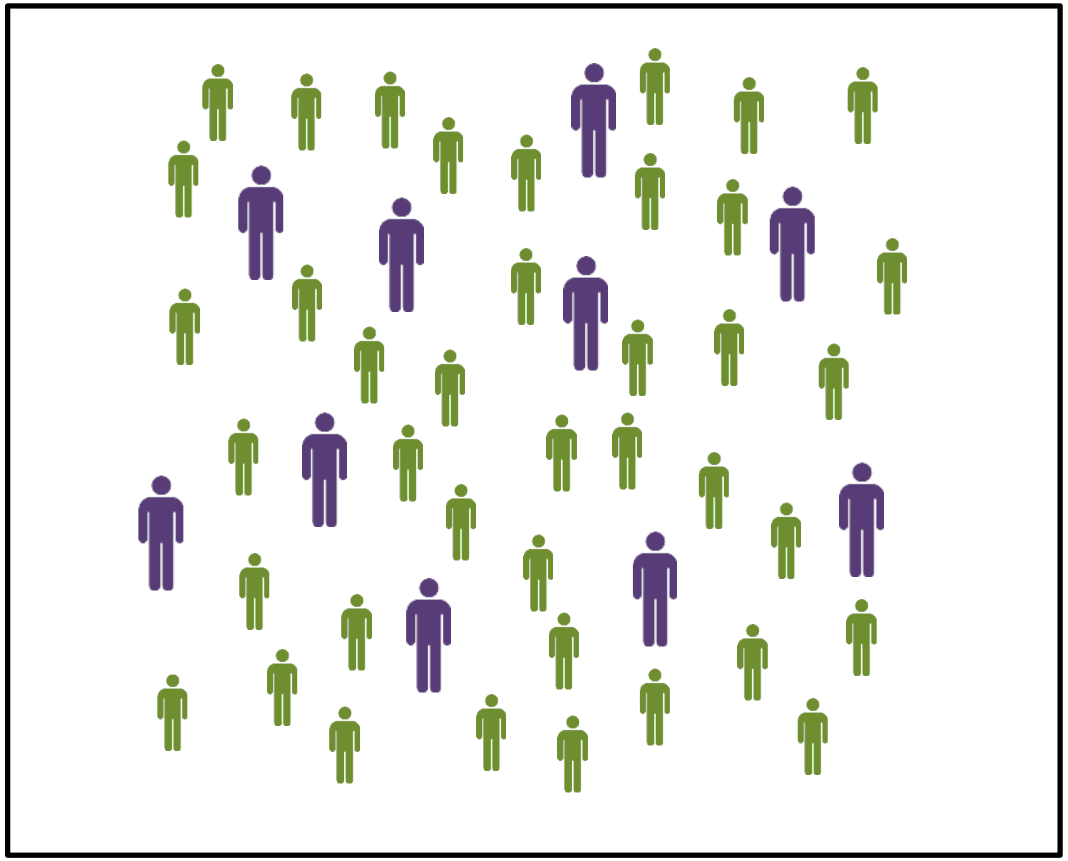
When you randomly sample groups of 100:

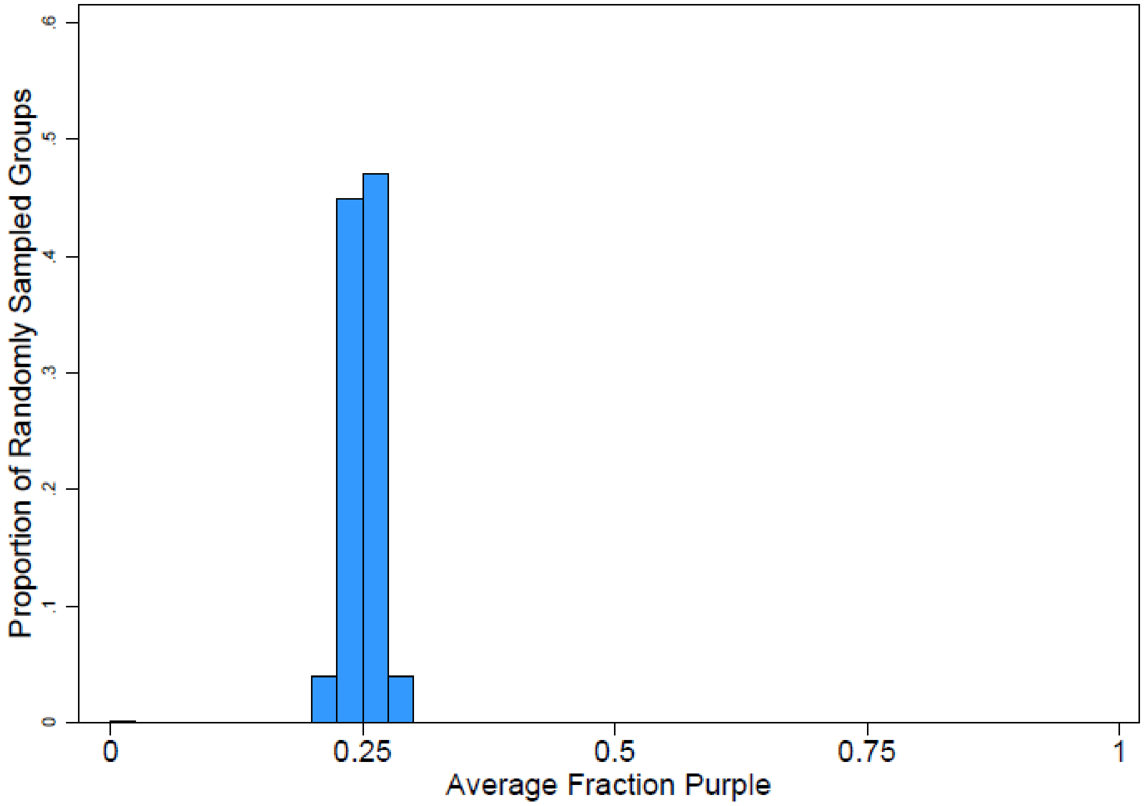




The Law of Large Numbers

When you randomly sample groups of 1,000:





Sampling Distribution

Aim of statistical inference is to **quantify the sample-to-sample variation**

This variation is captured by the **sampling distribution**

- e.g., Distribution of sample means from repeated sampling
- Not to be confused with sample distribution (= distribution of individual observations in the sample)

Measuring Variability

Variation in a distribution is captured by its **variance**:

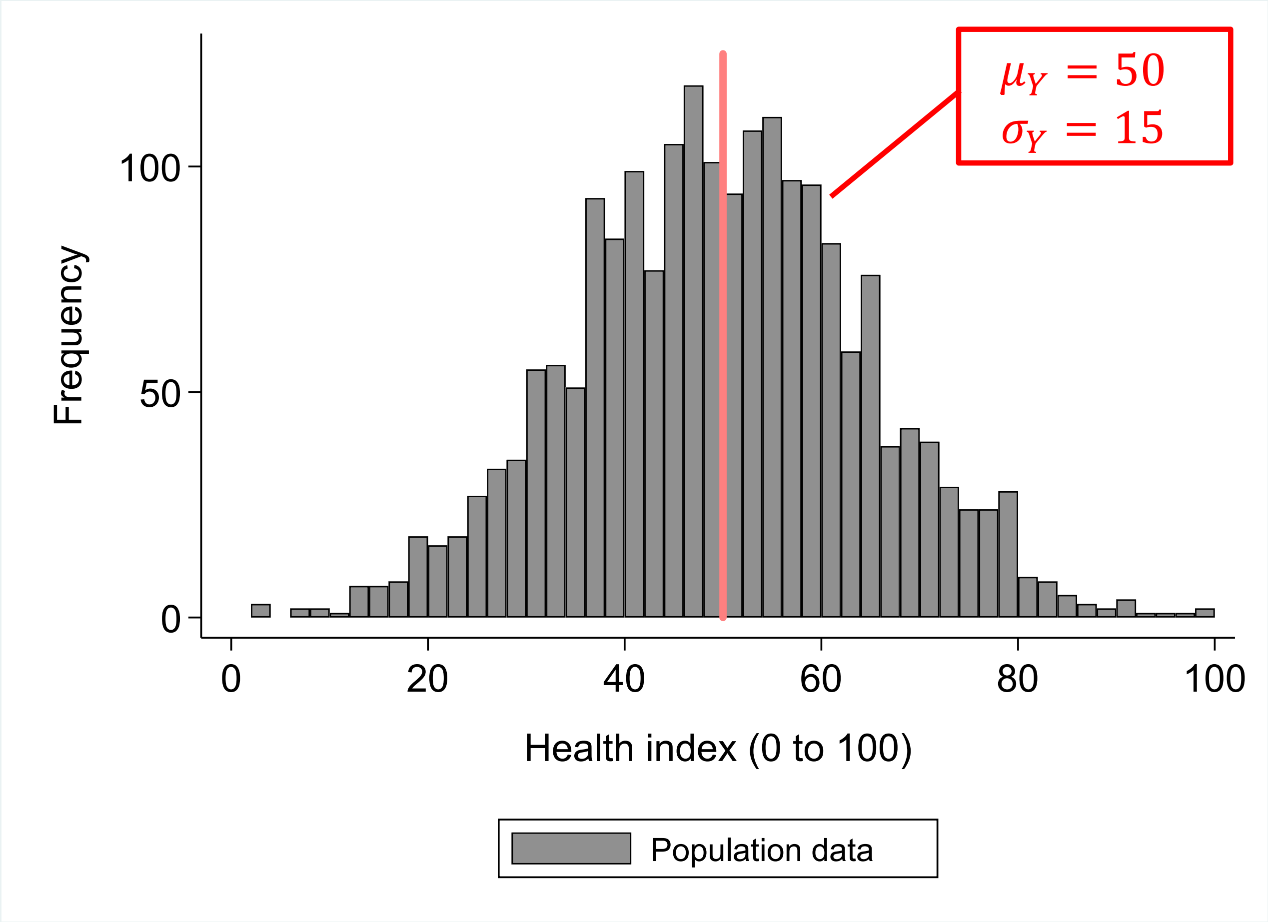
- Sample variance $\hat{V}$ of Y_i in a sample of size n: $\hat{V}(Y_i) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- Population variance V of Y_i as an expectation: $V(Y_i) = E[(Y_i - \bar{Y})^2]$

We typically work with the **standard deviation** (= square root of the variance):

$$\hat{\sigma}_Y = \sqrt{\hat{V}(Y_i)} \quad \text{and} \quad \sigma_Y = \sqrt{V(Y_i)}$$

Example – Student Health

Imagine we have health data on the entire population of 2,000 UMSI students:



Random Sampling Variation

How does a sample mean compare to the population mean?

- Take a random sample from this population ($n = 10$):

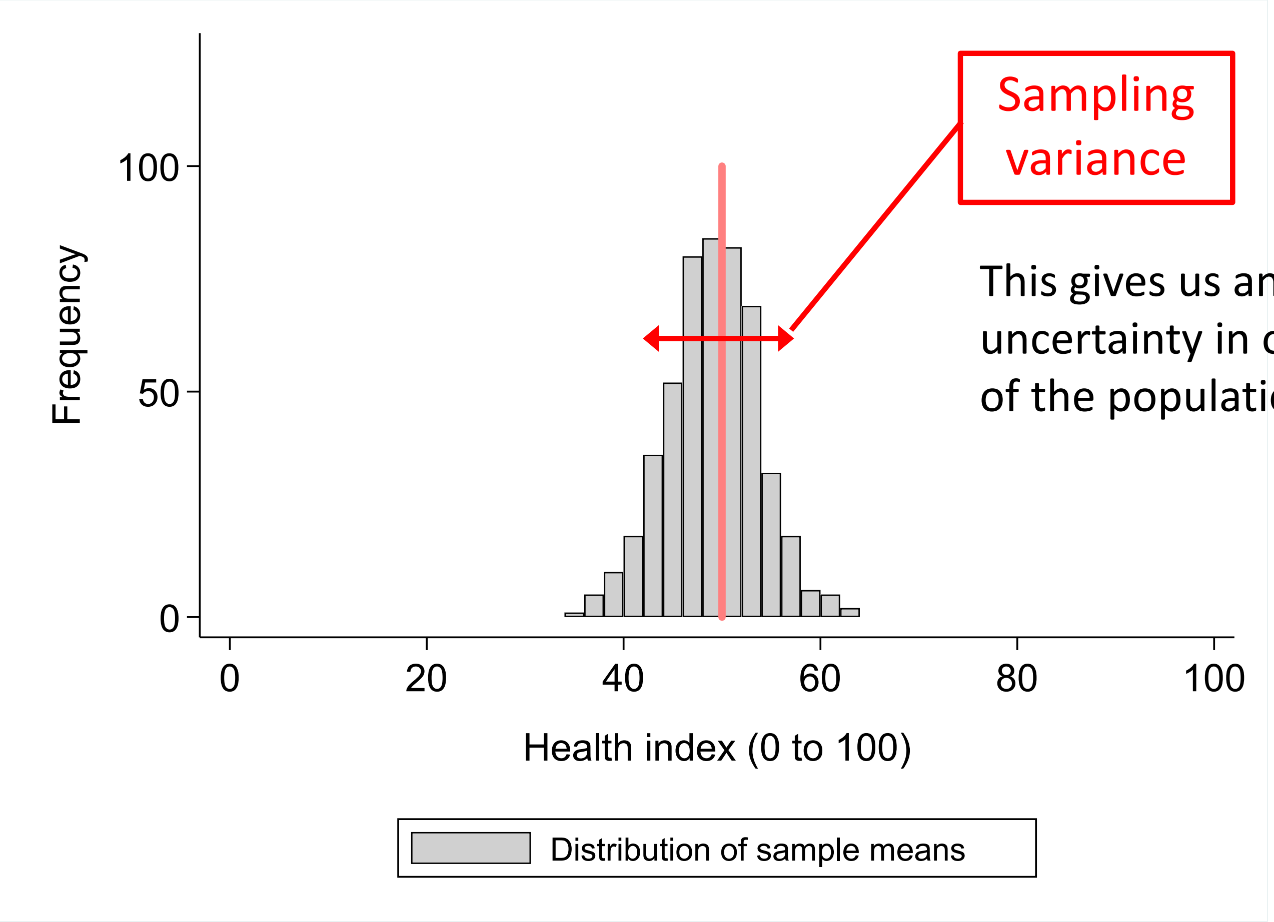
[66, 62, 35, 50, 40, 30, 54, 91, 34, 66]

- Is the sample mean close to the true population mean?
  - *Population* mean: 50
  - *Sample* mean: 52.8

- The difference is due to **random sampling variation**

Sampling Distribution

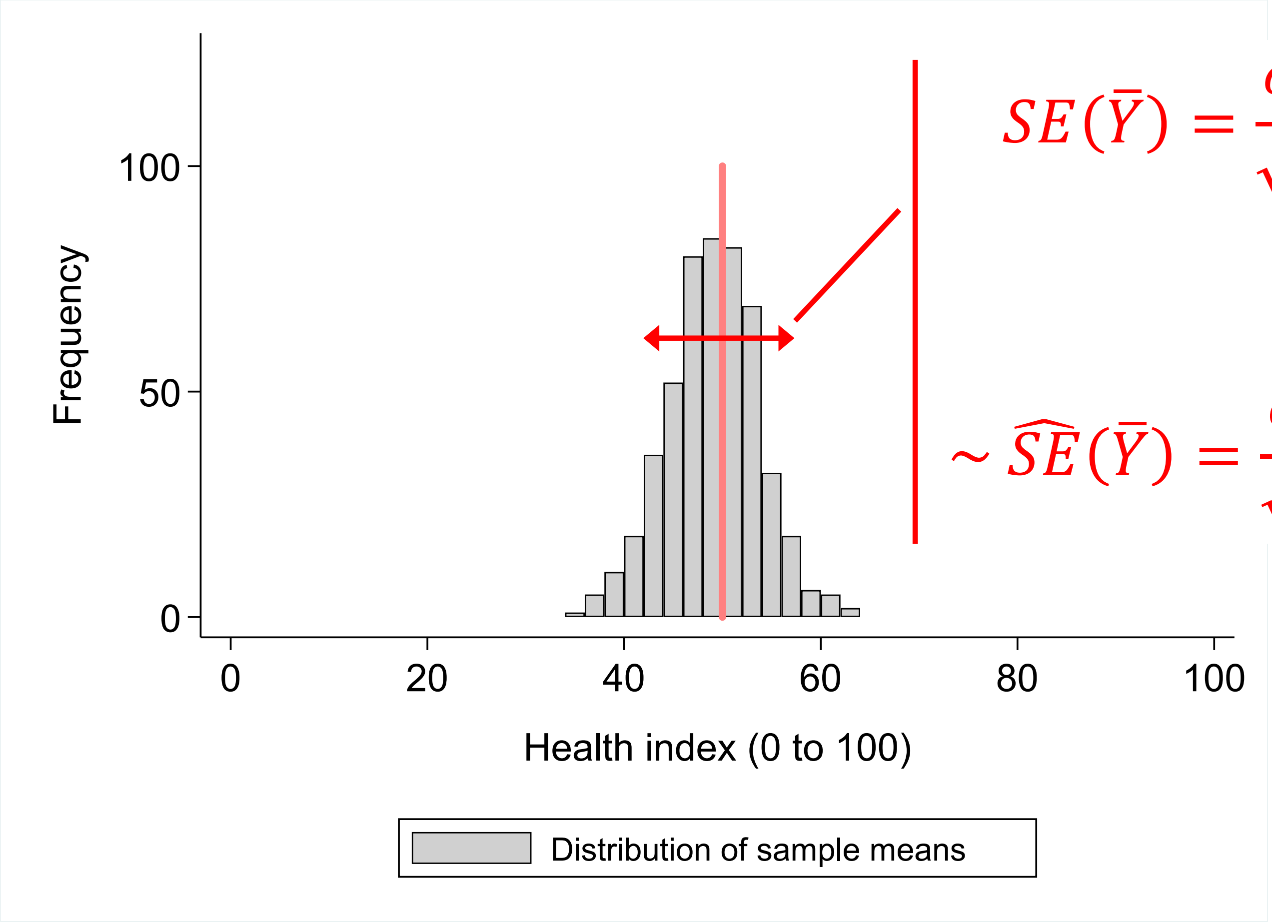
Distribution of 500 sample means (with $n = 10$ for each sample):



This gives us an idea of the uncertainty in our point estimate of the population mean.

Standard Error

Standard error (SE) = standard deviation of the sampling distribution



$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

$$\sim \widehat{SE}(\bar{Y}) = \frac{\hat{\sigma}_Y}{\sqrt{n}}$$

σ = population standard deviation

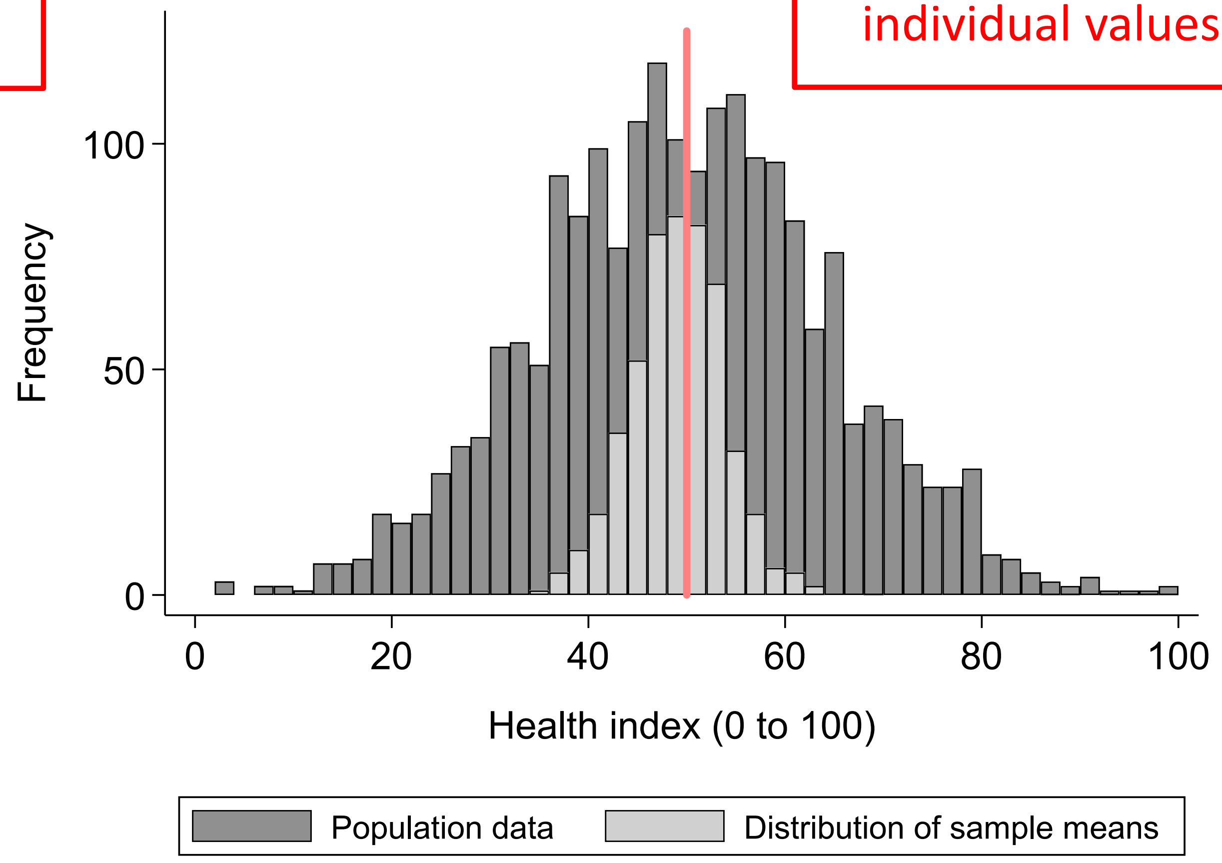
$\sqrt{n}$ = square of sample size

$\hat{\sigma}$ = estimate of the population standard deviation

Standard Error vs. Standard Deviation

Variability of estimates

Variability of individual values



Hypothesis Testing

Standard errors are used for **confidence intervals** (= range of plausible values for the parameter of interest)

Standard errors are also used for **hypothesis testing**:

- “What is the probability of getting a result as or more extreme than the sample result just by chance?”
- This probability is called *p-value*

The 3 Steps of Hypothesis Testing

1) Formulate the null hypothesis (and alternative hypothesis)

2) Calculate a test statistic (based on sample data)

3) Calculate a p-value using the probability distribution of the test statistic (and interpret the p-value considering the null hypothesis)

Example – Hypothesis Testing

Is the average health of UMSI students equal to 50?

- H_0: $E[Y_i] = \mu = 50$
- H_1: $\mu \neq 50$

Mean in a random sample of $n = 30$ students: $\bar{Y} = 57.6$

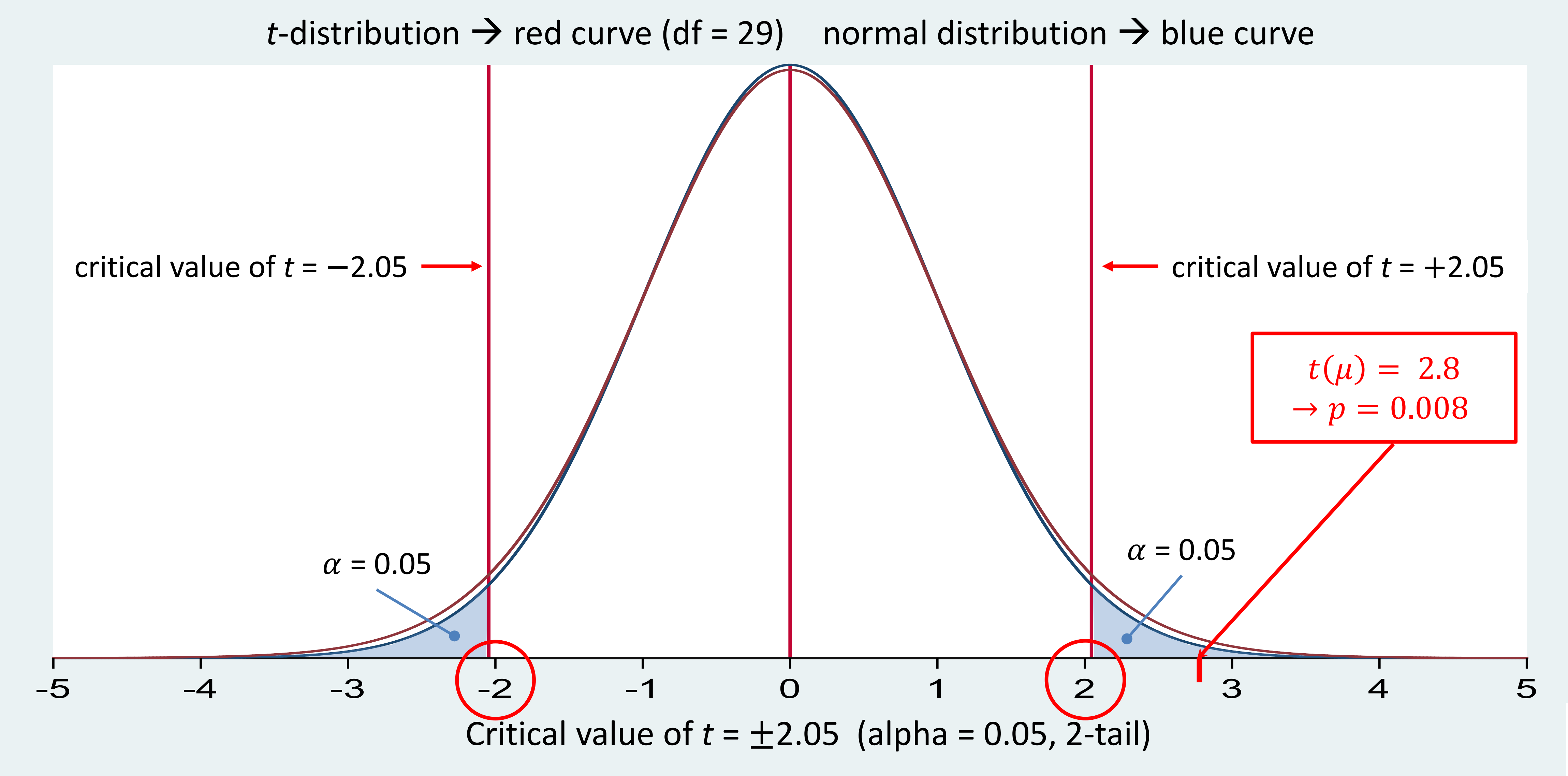
- $\widehat{SE}(\bar{Y}) = 2.7$

Given H_0 is true, what is the probability that we observe a value of 57.6 or more?

t-statistic:

$$t(\mu) = \frac{\bar{Y} - \mu}{\widehat{SE}(\bar{Y})} = \frac{57.6 - 50}{2.7} = 2.8$$

t-Distribution and p-Value



SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information

Randomized Experiments

Random Assignment Solves the Selection Problem

Selection bias occurs when we let people choose whether or not to get the treatment (e.g. health insurance):

$$E\left[Y_i^0 | D_i = 1\right] - E\left[Y_i^0 | D_i = 0\right] \neq 0$$

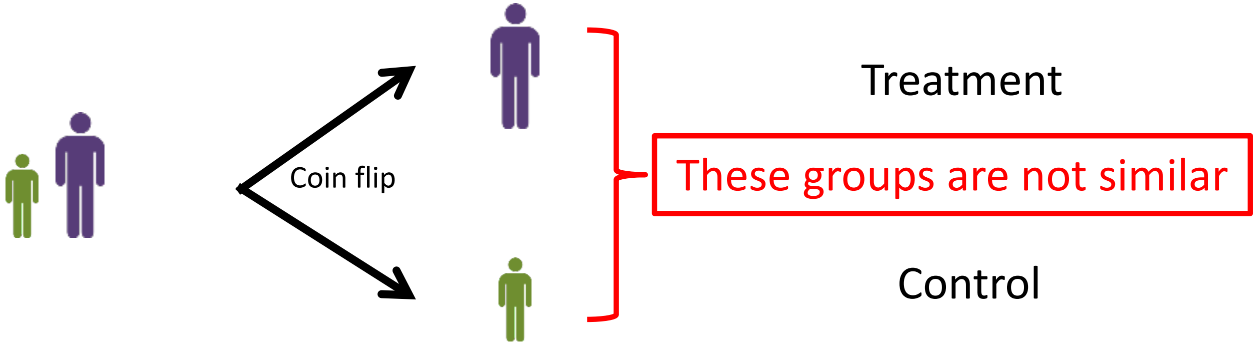
Suppose we *randomly* assign the treatment to individuals who are currently not treated (e.g. by a coin toss)

If the two groups are large enough, random assignment eliminates average *observed* and *unobserved* differences between the two groups:

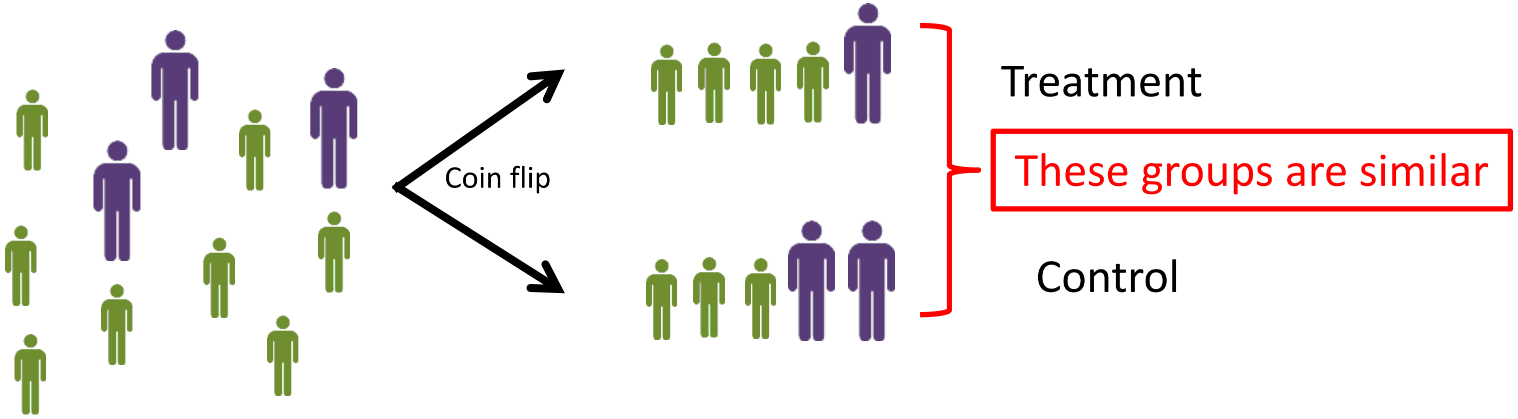
$$E\left[Y_i^0 | D_i = 1\right] = E\left[Y_i^0 | D_i = 0\right] = E[Y_i^0]$$

The Law of Large Numbers

When you randomly assign 2 individuals:



When you randomly assign *many* individuals:



Random Sampling vs. Random Assignment

Random Sampling	Random Assignment
• Law of large numbers (requires large enough samples)	• Law of large numbers (requires large enough samples)
• Sample average will be close to population average → *external validity* (or generalizability)	• Two randomly chosen groups will be similar → *internal validity*

Randomization Check

How do we know whether the randomization created two similar groups?

- It's impossible to know whether *all* variables are balanced between groups
- Best we can do is to check whether treatment and control groups are similar in terms of *observed* characteristics
- In practice, we usually use non-parametric tests to check for balance
  - Binary variables: Chi-squared test
  - Continuous variables: rank-sum test

Oregon's Health Insurance Experiment

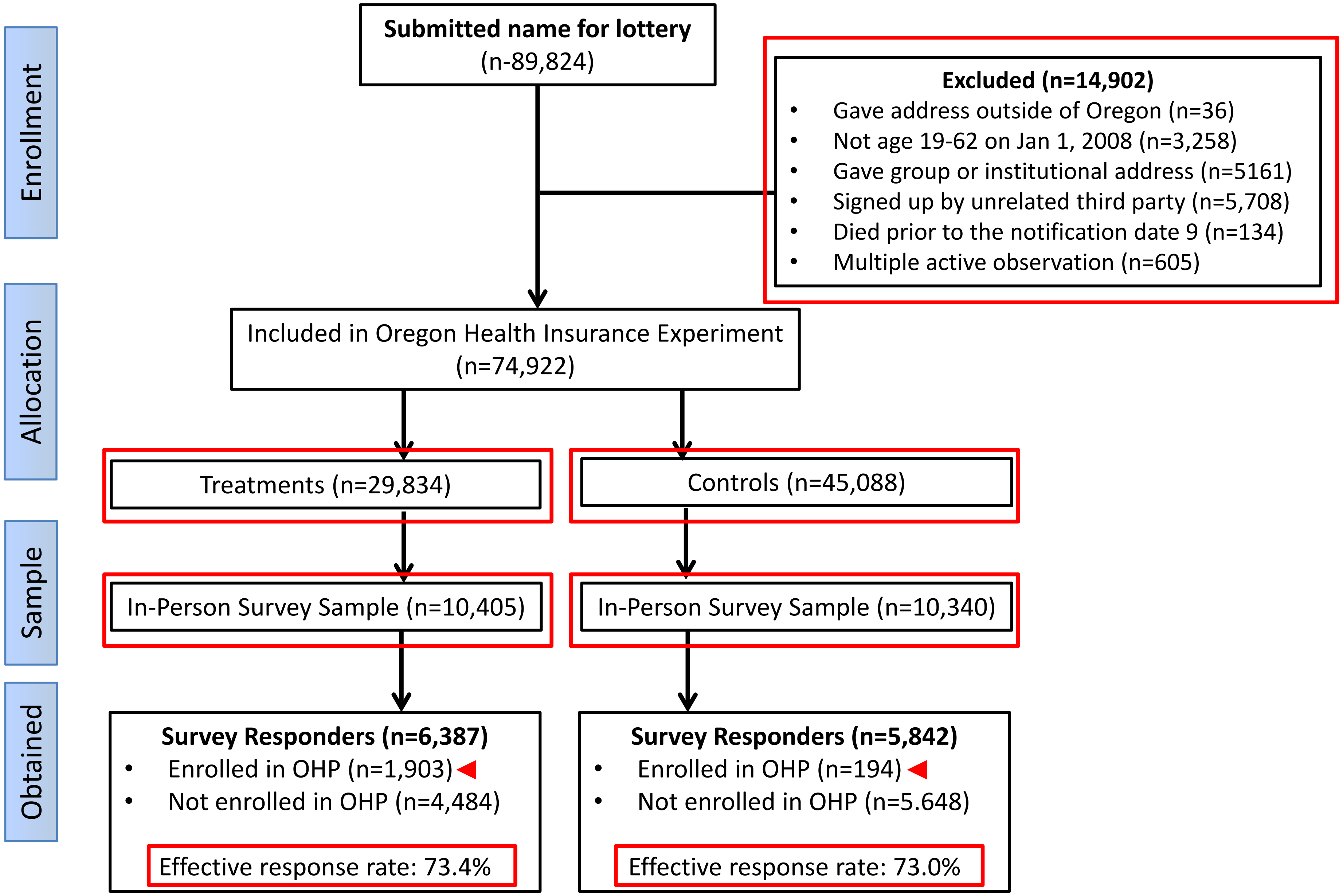
In 2008, Oregon expanded its Medicaid program to include low-income adults

- About 90,000 signed up for this program
- Due to budget constraints, only about 35,000 people were selected for the program
- State drew names from the waiting list by lottery
- Thus, the lottery created a randomized experiment to learn about the effects of Medicaid on health and financial outcomes

Finkelstein et al. (2012)

Problems with this experiment:

- ~30% of subsample did not complete the interview
- Only ~30% in treatment group were enrolled in Medicaid
- Some in control group were enrolled in Medicaid
- Decision to enroll likely not random



Noncompliance in Experiments

Some people did not comply with treatment assignment

- Some lottery winners were assigned to receive Medicaid but did not enroll
- Others qualified for regular Medicaid even without the lottery

Intention-to-treat analysis: redefine the treatment as "giving the *option* to receive treatment"

- ATE of receiving the option to enroll in Medicaid

Instrumental variable analysis (will be covered in a future module)

- Allows for estimating causal effect on having Medicaid
- Can only learn about compliers

In the Oregon health insurance experiment, lottery winners were about 26 percentage points more likely to be covered by Medicaid.

Medicaid ITT Effects on Health-Care Use

OHP effects on insurance coverage and health-care use

Outcome	Oregon Control mean (1)	Oregon Treatment effect (2)	Portland area Control mean (3)	Portland area Treatment effect (4)
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Lottery winners were about 10% *more* likely to use the emergency department

Medicaid ITT Effects on Health Outcomes

- Estimates are **intention-to-treat effects (ITT)**
- ITT effects likely underestimate true effects of Medicaid

OHP effects on health indicators and financial health

Outcome	Oregon: Control mean (1)	Oregon: Treatment effect (2)	Portland area: Control mean (3)	Portland area: Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

Lottery winners were 3.9%-points more likely to say their health is good or better

Almost no effects on objective measures of health

Medicaid ITT Effects on Medical Expenses

OHP effects on health indicators and financial health

Outcome	Oregon: Control mean (1)	Oregon: Treatment effect (2)	Portland area: Control mean (3)	Portland area: Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

- Medicaid program provided the safety net for which it was designed

Lottery winners are less likely to have large out-of-pocket medical expenses

Randomized Experiments

Randomized experiments are typically considered the “gold standard” because they eliminate selection bias on observables AND unobservables

Analysis of randomized experiments is simple (compare average outcomes across treatment and control groups)

If randomized experiments generate the most credible evidence and are easy to analyze, why don’t we always run experiments?

- Experiments can be costly and difficult to implement, and some may be unethical

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Credits:
Alain Cohn
Assistant Professor of Information