Research Project

Lecture 2

Comparison: From cause to effect

Adam.Booij@uva.nl

Trochim *et al.* Ch9.1/2;12.3d

2017-2018

UNIVERSITY OF AMSTERDAM

Aim of these lectures

Prepare you for conducting your own empirical study in block 6

Make you understand how to

1. "draw these lines": Ordinary Least Squares
   - An *effect* of interest in Micro, Macro, or Finance
2. interpret them given *sampling uncertainty*
   - t-test
3. make sure the interpretation makes sense *a priori*
   - empirical design
4. make sure the interpretation makes sense *a posteriori*
   - assessing design validity (covariate balance, F-test)

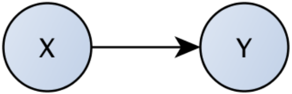
For a large part parallel to your Statistics course to make sure you *get practical understanding*!

Aim of this lecture

Make you understand

1. what is an “effect”
   - Potential outcomes
2. what is “sampling uncertainty” and how to deal with it
   - t-test

Causal effects



Empirical research

Most of empirical social science aims at estimating causal effects

- How will Y change if we change X_1, holding $W_1, \ldots, W_n$ constant?

Business and government often also interested in prediction

- What is the best prediction of Y_t given $Y_{t-1}, \ldots, Y_{t-k}$ and/or $X_1, \ldots, X_n$

We focus on causality

Causal effects

Examples:

- Effect of money supply on prices
- Effect of financial regulation on financial stability
- Effect of advertisement on sales
- Effect of job seek assistance on unemployement

Typically the (average) impact of a one unit "exogenous" change in "**treatment**" X_i on outcome variable of interest Y_i

Knowledge of causal effects important for economic policy & business decisions

Causal effects example: Jobs Seek Assistance

Job Seek Assistance (JSA) = Help for unemployed to find a job

- Professional assessment
- CV and motivation letters
- Search counseling
- Job support

JSA programs costly to implement

Government wants to know if effective (e.g. worth tax-payers money)

Large literature in labor economics

Causal effects and potential outcomes

How to think about the effect of JSA on unemployement?

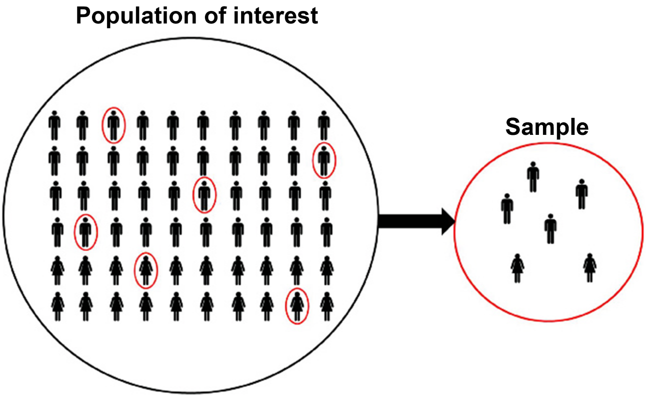
Use the "Rubin Causal Model" (RCM, JEduPsy'74)

Two "potential" outcomes:

- $Y_i(0)$: Does person i have a job in 8 months if no JSA
- $Y_i(1)$: Does person i have a job in 8 months if gets JSA

	Potential outcomes		Individual Treatment Effect
Person	$Y_i\,(0)$	$Y_i\,(1)$	
1	0	0	0
2	1	1	0
3	0	1	1
⋮	⋮	⋮	⋮
$N-1$	0	0	0
N	0	0	0

Sampling uncertainty



The population of potential outcomes

For the population of unemployed, the population "of interest", we want to know

$$ATE \equiv \underbrace{E\left(Y_i(1)\right)}_{\mu_1} - \underbrace{E\left(Y_i(0)\right)}_{\mu_0}$$

the **Average Treatment Effect** or **ATE**

		Potential outcomes		Treatment
Person		$Y_i(0)$	$Y_i(1)$	Effect
1		0	0	0
2		1	1	0
3		0	1	1
$\vdots$		$\vdots$	$\vdots$	$\vdots$
$N-1$		0	0	0
N		0	0	0
Average	$E()$	0.20	0.40	$\underbrace{\mathbf{0.20}}_{\textbf{ATE}}$

A sample of (potential) outcomes

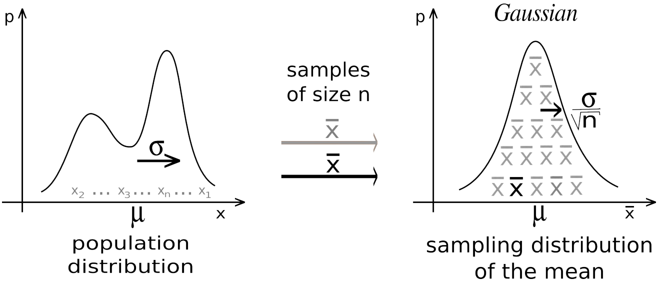
We only observe a sample, however,

Person	In Sample	X_i	Potential outcomes $Y_i(0)$	$Y_i(1)$	Treatment Effect
1	1	1	?	0	?
2	1	1	?	1	?
3	0	?	?	?	?
$\vdots$					
$N-1$	0	?	?	?	?
N	1	0	0	?	?
Average			0	0.5	0.5

Two complications

1. Practical: we can not *sample* every (unemployed) person $1, \ldots, N$
2. Fundamental: you get JSA ($X_i = 1$) or you don't ($X_i = 0$)
   - Never see both: the "*counterfactual* problem"

Dealing with sampling uncertainty: The central limit theorem



Random sampling and random assignment

The practical (sampling) and fundamental (counterfactual) problem prevent us to see the entire population

Say there are $N = 50,000$ unemployed people

Let us assume we have

1. Random sampling
2. Random assignment

Random sampling means all individuals have an equal chance to appear in the sample

Random assignment means all individuals have an equal chance to get JSA or not

Four example samples: draws

	Sample (1)		Sample (2)		Sample (3)		Sample (4)	
Person	In Sample	X_i	In Sample	X_i	In Sample	X_i	In Sample	X_i
1	0	?	1	1	0	?	1	1
2	1	1	0	?	1	1	0	?
3	1	0	0	?	1	0	1	1
$\vdots$								
$N-1$	0	?	0	?	1	1	1	0
N	1	0	1	0	0	?	1	1

Four example samples: observed outcomes

	Sample (1)		Sample (2)		Sample (3)		Sample (4)	
	Y_i		Y_i		Y_i		Y_i	
Person	$Y_i\,(0)$	$Y_i\,(1)$	$Y_i\,(0)$	$Y_i\,(1)$	$Y_i\,(0)$	$Y_i\,(1)$	$Y_i\,(0)$	$Y_i\,(1)$
1				0				0
2		1				1		
3	0				0			1
$\vdots$								
$N-1$	0					0	0	
N	0		0					0
Average	0	1	0	0	0	0.50	0	0.33

Four example samples: construction of the sample means

Consider sample 3

Assume the sample has $n = 1000$ with

- $n_0 = \sum_{i=1}^{n} (1 - X_i)$ people without JSA ("control group")
- $n_1 = \sum_{i=1}^{n} X_i$ people with JSA ("treatment group")

The sample means are defined as

- $\overline{Y}_0 = \frac{1}{n_0} \sum_{i:X_i=0} Y_i = \frac{1}{496} (0 + 0 + 0 + \ldots + 0) = \frac{0}{496} = 0$
- $\overline{Y}_1 = \frac{1}{n_1} \sum_{i:X_i=1} Y_i = \frac{1}{504} (0 + 1 + 1 + 0 \ldots + 0 + 1) = \frac{252}{504} = 0.5$

Many samples

	Sample mean	
	$X_i = 0$	$X_i = 1$
Draw	$\overline{Y}_0$	$\overline{Y}_1$
1	0	1
2	0	0
3	0	0.5
4	0	0.33
5	0.22	0.38
6	0.21	0.39
7	0.21	0.37
8	0.25	0.41
9	0.20	0.45
10	0.21	0.42
⋮		
10000	0.20	0.36
Average	?	?

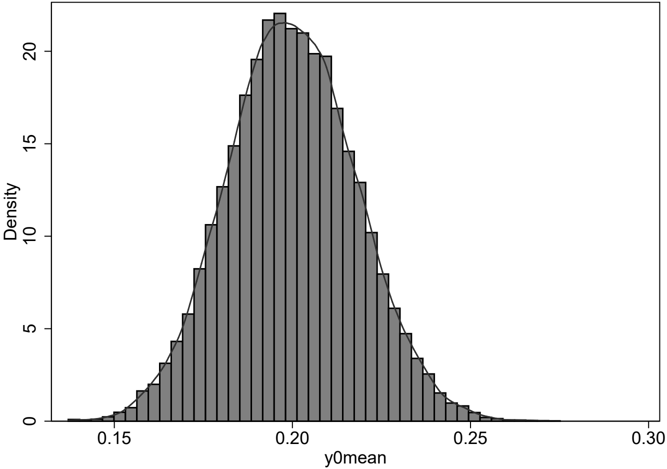
Many samples on average

	Sample mean	
	$X_i = 0$	$X_i = 1$
Draw	$\overline{Y}_0$	$\overline{Y}_1$
1	0	1
2	0	0
3	0	0.5
4	0	0.33
5	0.22	0.38
6	0.21	0.39
7	0.21	0.37
8	0.25	0.41
9	0.20	0.45
10	0.21	0.42
$\vdots$		
10000	0.20	0.36
Average	0.20	0.40

"on average correct": unbiased

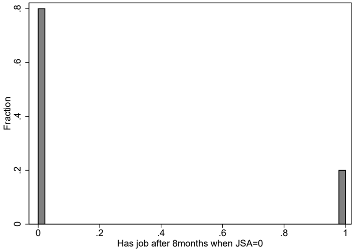
Many samples: distribution

Control group sample mean ($n = \pm 500$)



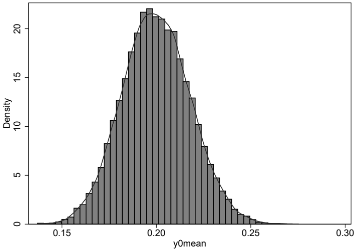
The Central Limit Theorem visual

Unit distribution



$Y_i(0)$

Sample mean distribution



$\overline{Y}_0$

The Central Limit Theorem formal

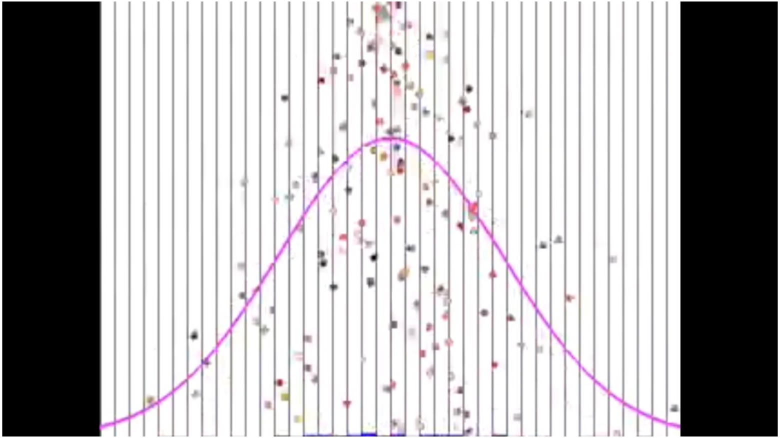
Central Limit Theorem

The sample average $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ calculated from *n independent and identically distributed draws* y_i, is, for a large enough sample:

- normally distributed with mean μ_y and standard deviation $\sigma_y/\sqrt{n}$ or
- $\overline{y} \overset{a}{\sim} N\left(\mu_y, \sigma_{\overline{y}}^2\right)$ with $\sigma_{\overline{y}}^2 = \sigma_y^2/n$

- This holds INDEPENDENT of the original distribution of y_i!!
- Often approximately true for n as low as 30 [rule of thumb]

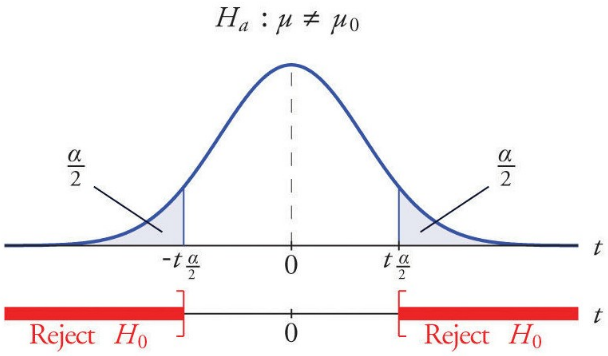
The Central Limit Theorem video



Central Limit Theorem video

Check out this more extensive video at home

Hypothesis testing: Using the central limit theorem to deal with sampling uncertainty



Descaling

The CLT says the sample mean $\overline{Y}_0 \sim N\left(\mu_0, \sigma_0^2/n\right)$, where $\sigma_0 = \sigma_{Y(0)}$

Because we do not observe σ_0 we can use a variant of the CLT using the so called

$$t - stat = \frac{\overline{Y}_0 - \mu_0}{SE\left(\overline{Y}_0\right)}$$

which uses $SE\left(\overline{Y}_0\right) = SD\left(Y_{0i}\right)/\sqrt{n_0}$ as an approximation for σ_0

In large samples

$$t - stat \sim N\left(0, 1\right) \text{ (std. normal)}$$

so we have removed the scale

Hypothesis testing: From sample to population

The $t - stat$ is very useful for doing *hypothesis testing*

Say our hypothesis is that without JSA, no one finds a job

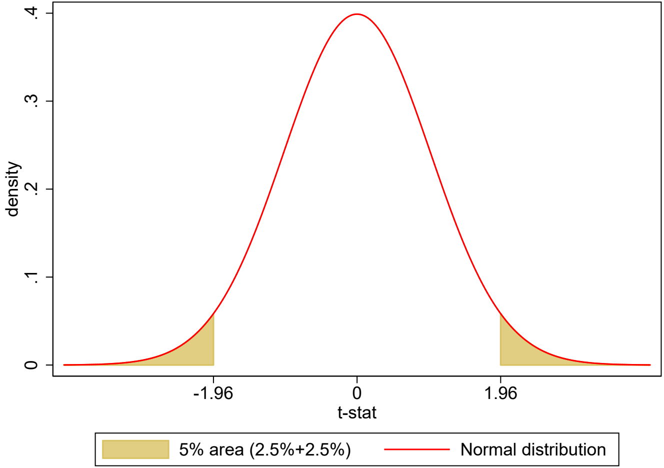
- $\mathcal{H}_0 : \mu_0 = 0$ vs $\mathcal{H}_1 : \mu_0 \neq 0$

A sample with $\overline{Y}_0 = 0.22$, $SD\left(Y_{0i}\right) = 0.42$ and $n_0 = 497$ would then, assuming $\mathcal{H}_0$ is true, give

$$t - stat = \frac{0.22 - 0}{0.42/\sqrt{497}} = 11.7$$

Is such a $t - stat$ likely?

The standard normal distribution



No! It is completely off the scale

Critical values of the std. normal distribution

When a $t - stat$ has a value that is unlikely to occur, we reject our hypothesis $\mathcal{H}_0$ and say there is evidence in favor of $\mathcal{H}_1$

More formally when

$$|t - stat| > t_{\alpha/2}^{crit} = 1.96$$

we reject $\mathcal{H}_0$, using the most commonly used "significance level" $\alpha = 0.05 = 5\%$

The significance level is the probability of rejecting $\mathcal{H}_0$ while in fact it is true; the Type I error

We take the absolute value because we do a two-sided test (more on that in the next lecture)

The 95% confidence interval

We have a sample with $\overline{Y}_0 = 0.22$, $SD\left(Y_{0i}\right) = 0.42$ and $n_0 = 497$

The $SE\left(\overline{Y}_0\right) = SD\left(Y_{0i}\right)/\sqrt{n_0} = 0.42/\sqrt{497} = 0.019$

What hypothesized values of μ_0 would be rejected by the data?

Hypotheses that are more than $1.96 \cdot SE\left(\overline{Y}_0\right)$ away from $\overline{Y}_0 = 0.22$ will be rejected

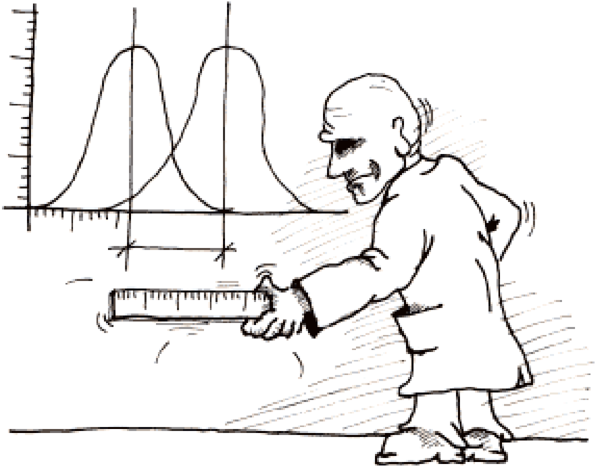
This is the 95% confidence interval

$$\left[\overline{Y}_0 - 1.96 \cdot SE\left(\overline{Y}_0\right); \overline{Y}_0 + 1.96 \cdot SE\left(\overline{Y}_0\right)\right]$$

$$[0.183; 0.257]$$

Will 0 be rejected? Will 0.2 be rejected?

Comparison: Testing the difference of two sample means



Mean comparison

The standard

$$t - stat = \frac{\overline{Y}_0 - \mu_0}{SE\left(\overline{Y}_0\right)}$$

is useful for testing the population mean of a variable

Most often, however, we will be interested in comparing the mean of a variable between two independent groups

In our context: does JSA have an effect?

- $\mathcal{H}_0 : \mu_0 = \mu_1$ vs $\mathcal{H}_1 : \mu_0 \neq \mu_1$

Mean comparison t-stat

We can adapt the $t - stat$ to compare two groups by

$$t - stat = \frac{\overline{Y}_1 - \overline{Y}_0 - (\mu_1 - \mu_0)}{SE\left(\overline{Y}_1 - \overline{Y}_0\right)}$$

where $SE\left(\overline{Y}_1 - \overline{Y}_0\right) = \sqrt{\frac{SD^2(Y_{0i})}{n_0} + \frac{SD^2(Y_{1i})}{n_1}}$ is the standard error of the difference

Mean comparison t-stat example

With (independent) samples

- JSA=0: $\overline{Y}_0 = 0.22$; $SD\left(Y_{0i}\right) = 0.42; n_0 = 497$
- JSA=1: $\overline{Y}_1 = 0.38$; $SD\left(Y_{1i}\right) = 0.49; n_1 = 503$

our best guess is that the effect of JSA is

$$\hat{ATE} = \overline{Y}_1 - \overline{Y}_0 = 0.38 - 0.22 = 0.16$$

Is this significant evidence of an effect of JSA on finding a job?

Using $SE\left(\overline{Y}_1 - \overline{Y}_0\right) = \sqrt{\frac{0.42^2}{497} + \frac{0.49^2}{503}} = 0.029$ we calculate the

$$t - stat = \frac{0.16 - 0}{0.029} = 5.52$$

$|t - stat| > 1.96$ so we reject $\mathcal{H}_0 : \mu_0 = \mu_1$ in favor of $\mathcal{H}_1 : \mu_0 \neq \mu_1$

Mean comparison t-stat example 95% CI

What hypothesized values of $\mu_1 - \mu_0$ would be rejected by the data?

Hypotheses that are more than $1.96 \cdot SE\left(\overline{Y}_1 - \overline{Y}_0\right)$ away from $\hat{ATE} = 0.16$ will be rejected

This is the 95% confidence interval

$$\left[\hat{ATE} - 1.96 \cdot SE\left(\overline{Y}_1 - \overline{Y}_0\right); \hat{ATE} + 1.96 \cdot SE\left(\overline{Y}_1 - \overline{Y}_0\right)\right]$$

$$[0.103; 0.217]$$

Will 0.1 be rejected? Will 0.2 be rejected?

Is this good news for JSA programs?

Remark on the t-stat

It may be confusing that we use the $t-stat$ but say it is *normally distributed*

When the original (unit) variables are normally distributed, we know that the $t-stat$ is *exactly* t_n distributed for any sample size n

For large n, t_n becomes $N(0,1)$

In practice the unit distributions will (almost) never be normal

We thus always use the normal distribution as an approximation of the small sample distribution

For samples above $n > 120$ this approximation usually works well

Summary

What did we discuss today?

- Causal effects
- Sampling uncertainty due to
  - small sample
  - counterfactual problem
- Hypothesis testing: t-test
  - single mean
  - mean comparison

Next: Comparison of two groups when assignment is not random