Potential outcomes and the experimental ideal

Mauricio Romero

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes and the experimental ideal

Our goal

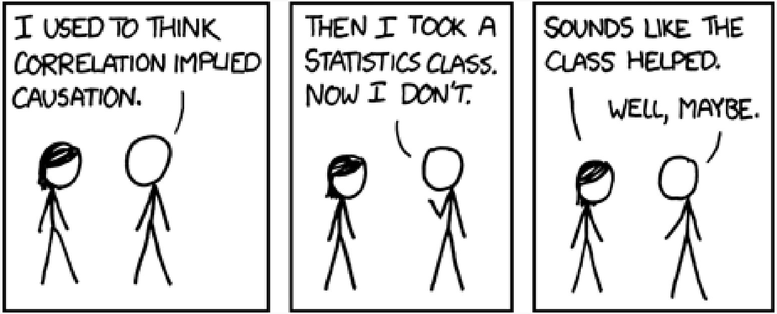
Potential outcomes

Correlation, causation and false counterfactuals

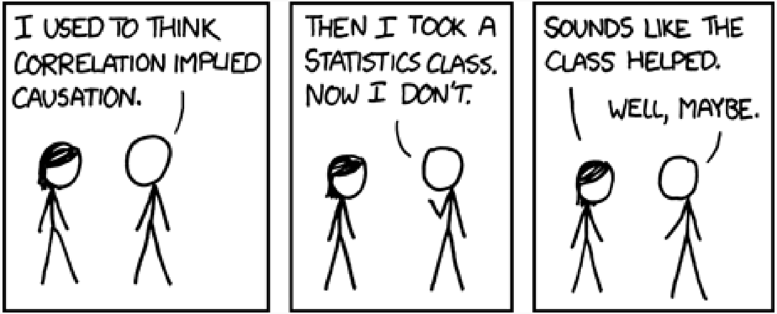
Experimental ideal

Randomization and selection bias

Randomization inference


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://m.xkcd.com/552/


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

`https://m.xkcd.com/552/`

Disclaimer: This class draws heavily on material from Scott Cunningham's mixtape, Nick Huntington-Klein's classes, Angrist and Pischke "Mostly Harmless Econometrics" and "Mastering Metrics", Stock and Watson's "Introduction to Econometrics", and other places.

Cause and effect

- We are interested in the relationship between “treatment” and some outcome
    - Treatment: Some drug; Outcome: health status
    - Treatment: Attending school; Outcome: wages
    - Treatment: Waking-up early; Outcome: learning
    - Treatment: Drinking alcohol; Outcome: child development
    - Treatment: Legalizing weed; Outcome: violence

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}

Potential outcomes

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}
- The observed outcome

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i}) T_i \end{aligned}$$

Potential outcomes

- A treatment (T) induces two "potential outcomes" for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}
- The observed outcome

$$
\begin{aligned}
Y_i &= \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases} \\
&= Y_{0i} + (Y_{1i} - Y_{0i}) T_i
\end{aligned}
$$

- The impact for any individual is $\delta_i = Y_{1i} - Y_{0i}$

Potential outcomes

- A treatment (T) induces two “potential outcomes” for individual i
  - The untreated outcome Y_{0i}
  - The treated outcome Y_{1i}
- The observed outcome

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})T_i \end{aligned}$$

- The impact for any individual is $\delta_i = Y_{1i} - Y_{0i}$
- Fundamental problem: **Never observe both potential outcomes for the same individual**

We can't just compared treated/untreated individuals

- We observe $Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\delta_i = \text{impact}} T_i$
- If we compare the outcomes of treated and untreated individuals:

$$\underbrace{\mathbb{E}(Y_i | T_i = 1) - \mathbb{E}(Y_i | T_i = 0)}_{\text{Observed difference}} \quad =$$

We can't just compared treated/untreated individuals

- We observe $Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\delta_i = \text{impact}} T_i$
- If we compare the outcomes of treated and untreated individuals:

$$\begin{aligned}\underbrace{\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)}_{\text{Observed difference}} \quad = \quad & \mathbb{E}(Y_{1i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 1) + \\ & \mathbb{E}(Y_{0i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 0)\end{aligned}$$

We can't just compared treated/untreated individuals

- We observe $Y_i = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\delta_i = \text{impact}} T_i$

- If we compare the outcomes of treated and untreated individuals:

$$\begin{aligned}
\underbrace{\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)}_{\text{Observed difference}} &= \mathbb{E}(Y_{1i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 1) + \\
&\quad \mathbb{E}(Y_{0i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 0) \\
&= \underbrace{\mathbb{E}(Y_{1i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 1)}_{\text{average treatment effect on the treated}} + \\
&\quad \underbrace{\mathbb{E}(Y_{0i}|T_i = 1) - \mathbb{E}(Y_{0i}|T_i = 0)}_{\text{selection bias}}
\end{aligned}$$

An example with simulations – Roy Model

- Assume people's potential outcomes are distributed:

$$\begin{aligned} Y_{0i} &\sim N(0,1) \\ Y_{1i} &\sim N(0.2,1) \end{aligned}$$

- This implies the average treatment effect (ATE) is 0.2
- While the potential outcome if treated is greater **on average** ($\mathbb{E}Y_{1i} > \mathbb{E}Y_{0i}$)
- For any given individual $Y_{1i} \underbrace{\overset{\leq}{\underset{>}{=}}}_{?} Y_{0i}$

An example with simulations – Roy Model

So what if people only get treatment if $Y_{1i} > Y_{0i}$

```
## Roy model and selection bias
N=100000 #number of observations
Y0 <-  rnorm(n=N, mean=0, sd=1)# control potential outcome
Y1 <-  rnorm(n=N, mean=0.2, sd=1)# treatment potential outcome
mean(Y1)-mean(Y0)
#You only get treatment if Y1>Y0
Treatment=(Y1>Y0)
#What we observe
Y=Y1*Treatment+Y0*(1-Treatment)
#Lets look at the difference in means across treated/untreated individuals
mean(Y[Treatment==1])-mean(Y[Treatment==0])
```

An example with simulations – Roy Model

- While the treatment effect was 0.2
- ...the observe difference across treated/untreated is ≈ 0.067 (pretty far off)
- Under our assumptions, you could work the math:

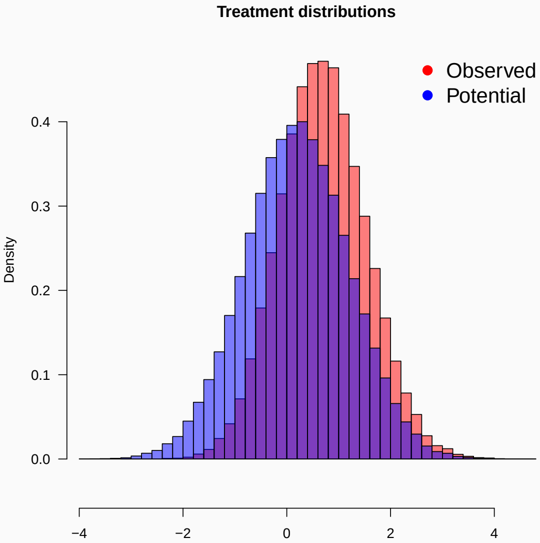
$$\begin{aligned}
\underbrace{\mathbb{E}(Y_i|T_i=1)-\mathbb{E}(Y_i|T_i=0)}_{\text{Observed difference}} &= \underbrace{\mathbb{E}(Y_{1i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=1)}_{\text{average treatment effect on the treated}}+ \\
&\quad \underbrace{\mathbb{E}(Y_{0i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=0)}_{\text{selection bias}} \\
&= \mathbb{E}(Y_{1i}|Y_{1i}>Y_{0i})-\mathbb{E}(Y_{0i}|Y_{1i}>Y_{0i})+ \\
&\quad \mathbb{E}(Y_{0i}|Y_{1i}>Y_{0i})-\mathbb{E}(Y_{0i}|Y_{1i}<Y_{0i})
\end{aligned}$$

- You learned this in your stats class: you can calculate the treatment effect under our assumptions and the selection bias!

Difference between distribution of “potential outcomes” and the observe outcomes (conditional on treatment)

```r
#distribution of observed and potential outcomes of the treated
hist(Y[Treatment==1],col=rgb(1,0,0,0.5),breaks=50,freq=F,
       las=1,xlab="",xlim=range(c(Y,Y1)),axes=F,
       main="Treatment distributions")
axis(side=2,las=1)         ## add default y-axis (ticks+labels)
axis(side=1,las=1,line=2)
hist(Y1,add=T,col=rgb(0,0,1,0.5),breaks=50,freq=F)
legend("topright",c("Observed","Potential"),
       col=c(rgb(1,0,0,1),rgb(0,0,1,1)),
       pch=19,bty="n",cex=1.5)
```

Bias in treated distribution



Some important definitions

- The average treatment effect (ATE)

$$\mathbb{E}[\delta_i] \quad = \quad \mathbb{E}[Y_{i1} - Y_{i0}]$$

- The average treatment effect on the treated (ATT) is the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} \mathbb{E}[\delta_i | T_i = 1] &= \mathbb{E}[Y_{i1} - Y_{i0} | T_i = 1] \\ &= \mathbb{E}[Y_{i1} | T_i = 1] - \mathbb{E}[Y_{i0} | T_i = 1] \end{aligned}$$

- The average treatment effect on the untreated (ATU) is the average treatment effect conditional on being untreated:

$$\begin{aligned} \mathbb{E}[\delta_i | T_i = 0] &= \mathbb{E}[Y_{i1} - Y_{i0} | T_i = 0] \\ &= \mathbb{E}[Y_{i1} | T_i = 0] - \mathbb{E}[Y_{i0} | T_i = 0] \end{aligned}$$

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

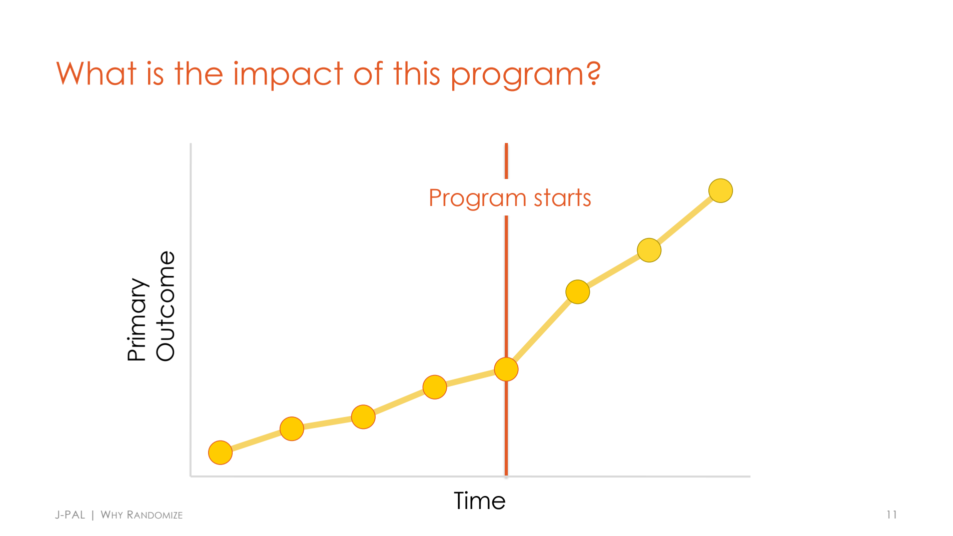
Randomization and selection bias

Randomization inference

Two types of false counterfactuals:

- Participant vs. Non-Participant comparisons (we just saw why this is problematic)
- Pre-treatment vs. Post-treatment comparisons (why?)

Pre-treatment vs. Post-treatment comparisons (from J-PAL slides on why randomize)

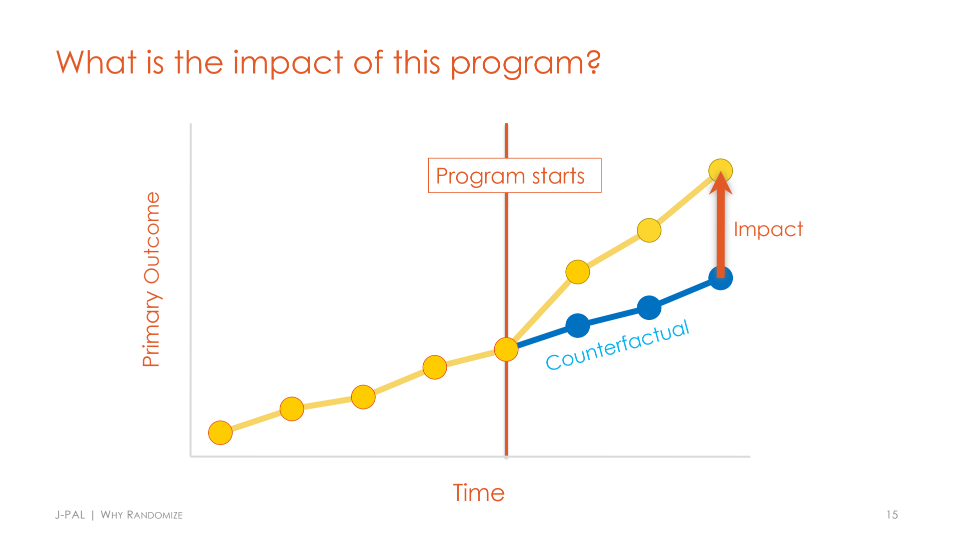


What is the impact of this program?

J-PAL | Why Randomize

11

Pre-treatment vs. Post-treatment comparisons (from J-PAL slides on why randomize)

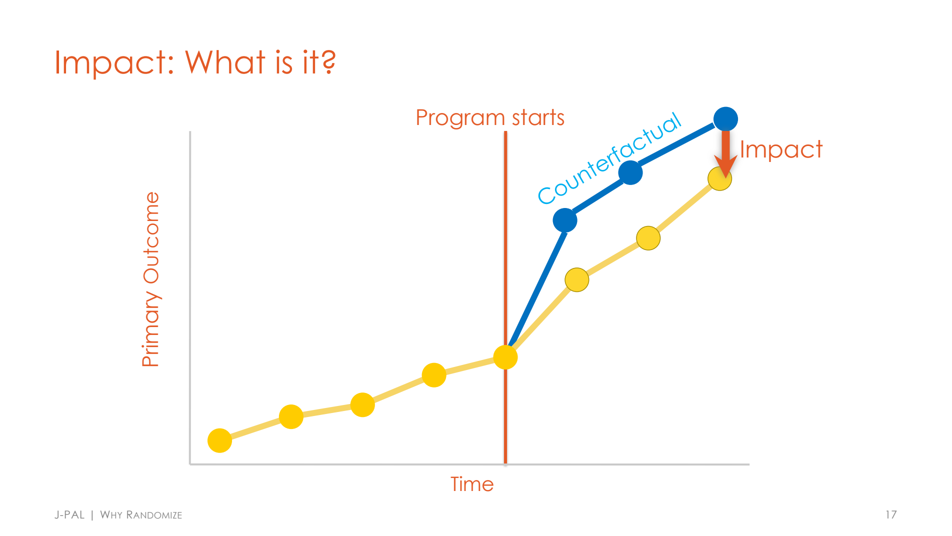


What is the impact of this program?

J-PAL | Why Randomize

15

Pre-treatment vs. Post-treatment comparisons (from J-PAL slides on why randomize)

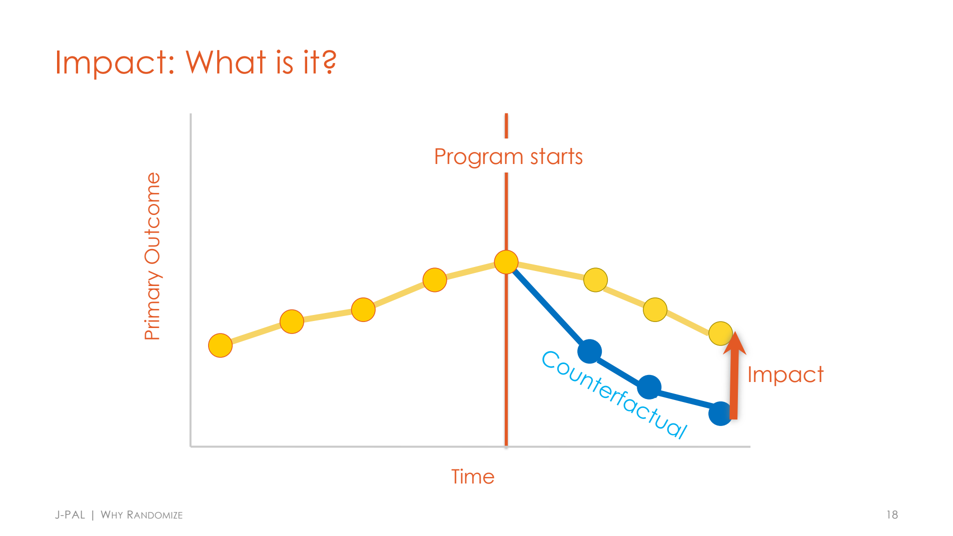


Impact: What is it?

J-PAL | Why Randomize

17

Pre-treatment vs. Post-treatment comparisons (from J-PAL slides on why randomize)



Impact: What is it?

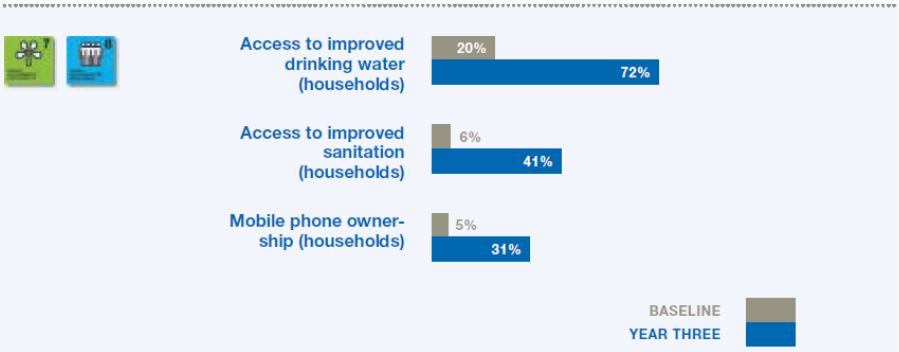
J-PAL | WHY RANDOMIZE

18

Millennium development villages

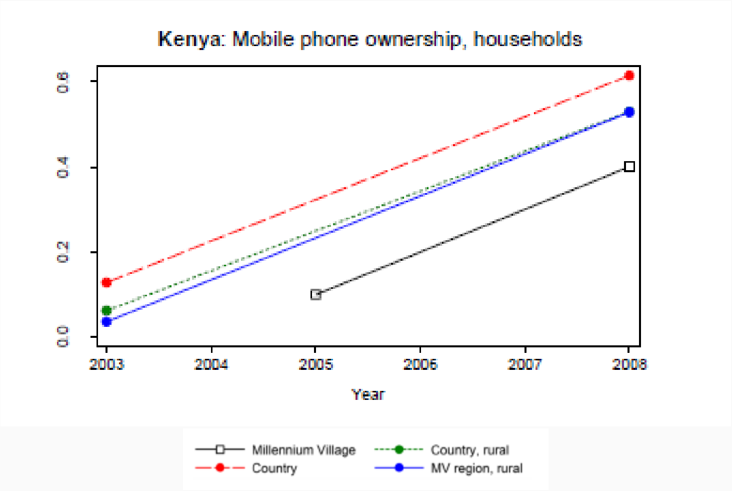
- Famous example of pre/post comparisons
- First evaluation relied on data on pre-treatment and post-treatment outcomes in Bar Sauri, Kenya
- On most outcomes people living in the MDV looked better after a 3-5 years
- But Clemens-Demombynes (2010) show the country as a whole looked better too
- See here for more papers:
  - `https://www.pnas.org/content/104/43/16775`
  - `https://www.cgdev.org/publication/when-does-rigorous-impact-evaluation-make-difference-case-millennium-villages-working`
  - `https://www.sciencedirect.com/science/article/pii/S2214109X18300652`

Millennium development villages: Pre-post comparisons



Harvests of Development in Rural Africa: The Millennium Villages After Three Years

Millennium development villages: Clemens-Demombynes (2010)



Correlation and causality are very different concepts

- Causal question: "If I go to the hospital (T), will my health (Y) improve?"

- Correlation question:

$$\frac{Cov(T, Y)}{\sigma_T \sigma_Y}$$

- These are not the same thing

Correlation $\neq$ Causality: Reverse causality

- If two variables (A and B) are correlated, does A cause B, or the other way around (or neither)?

Correlation $\neq$ Causality: Reverse causality

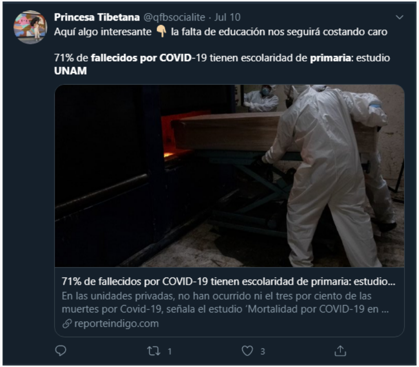
- If two variables (A and B) are correlated, does A cause B, or the other way around (or neither)?
- In the Middle Ages believed that lice were beneficial to your health
- There would rarely be any lice on sick people
- Thus (they argued) people got sick because the lice left
- Real reason: lice are extremely sensitive to body temperature. When you get a fever lice will look for another host
- Thus, you have no lice because you are sick, not the other way around
- See `https://blogs.scientificamerican.com/guest-blog/of-lice-and-men-an-itchy-history/`

Correlation $\neq$ Causality: A third factor

- If two variables (A and B) are correlated, a third factor (C) may cause both

Correlation ≠ Causality: A third factor

- If two variables (A and B) are correlated, a third factor (C) may cause both



Princesa Tibetana @qfbsocialite · Jul 10

Aquí algo interesante 👇 la falta de educación nos seguirá costando caro

71% de **fallecidos por COVID**-19 tienen escolaridad de **primaria**: estudio **UNAM**

71% de fallecidos por COVID-19 tienen escolaridad de primaria: estudio...

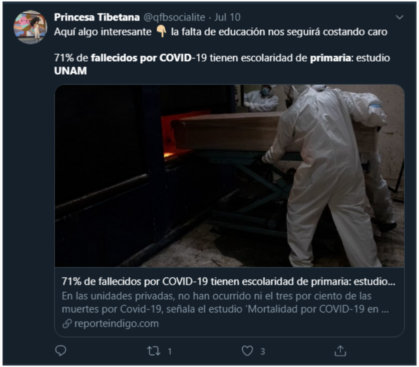
En las unidades privadas, no han ocurrido ni el tres por ciento de las muertes por Covid-19, señala el estudio 'Mortalidad por COVID-19 en ...

🔗 reporteindigo.com

https://twitter.com/qfbsocialite/status/1281632279308066818?s=20

Correlation $\neq$ Causality: A third factor

- If two variables (A and B) are correlated, a third factor (C) may cause both



Princesa Tibetana @qfbsocialite · Jul 10

Aquí algo interesante 👇 la falta de educación nos seguirá costando caro

71% de **fallecidos por COVID**-19 tienen escolaridad de **primaria**: estudio **UNAM**

71% de fallecidos por COVID-19 tienen escolaridad de primaria: estudio...

En las unidades privadas, no han ocurrido ni el tres por ciento de las muertes por Covid-19, señala el estudio 'Mortalidad por COVID-19 en ...

🔗 reporteindigo.com

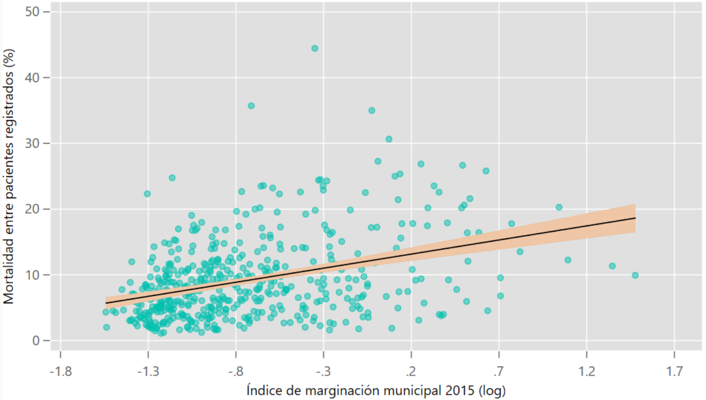
https://twitter.com/qfbsocialite/status/1281632279308066818?s=20

- Likely, poverty is driving both low educational attainment and COVID mortality

Correlation ≠ Causality: A third factor

Desigualdad en la pandemia: Marginación municipal vs. mortalidad entre pacientes registrados
Municipios con al menos 10 muertes (807 municipios)

(Total acumulado desde el primer caso en México hasta el 25 de julio. Cada punto representa un municipio).
Predicción lineal con intervalo de confianza de 95%



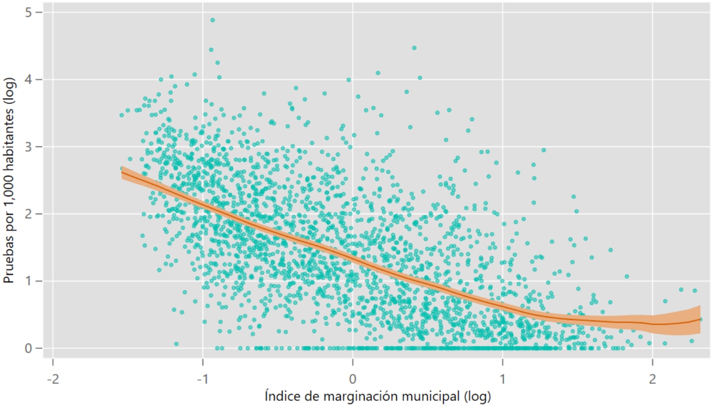
Fuente: Mariano Sánchez Talanquer, "Inequality in the Pandemic: Evidence from Mexican municipalities", working paper.

`https://twitter.com/mstalanquer/status/1287478515135385601?s=20`

Correlation $\neq$ Causality: A third factor

Desigualdad en la pandemia: Marginación municipal e incidencia de pruebas por 1,000 hab.

(Total acumulado desde la primera prueba en México hasta el 25 de julio. Cada punto representa un municipio.
La línea muestra un polinomio local con intervalo de confianza de 95%)



Fuente: Mariano Sánchez Talanquer, "Inequality in the Pandemic: Evidence from Mexican municipalities", working paper.

https://twitter.com/mstalanquer/status/1287478515135385601?s=20

Correlation $\neq$ Causality: Bidirectional causation

- If two variables (A and B) are correlated, A may cause B and B may cause A

- This includes a lot of relationships in ecology (and evolution)

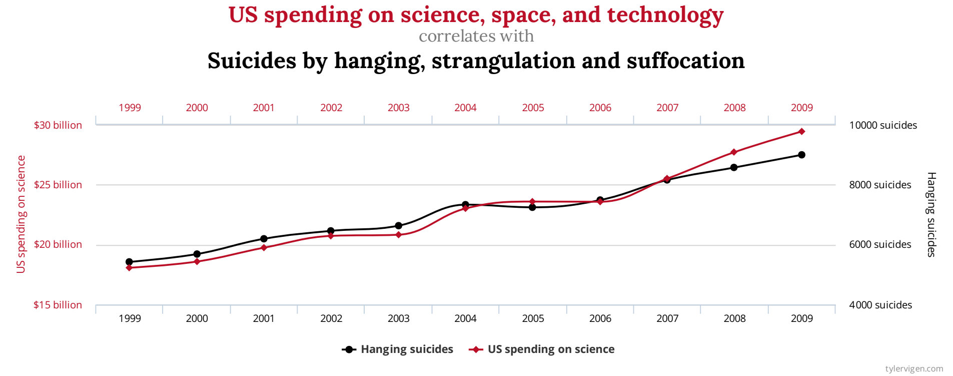
Correlation $\neq$ Causality: Bidirectional causation

- If two variables (A and B) are correlated, A may cause B and B may cause A
- This includes a lot of relationships in ecology (and evolution)
- Another example: Depression and drugs
  - Depression can lead to drug abuse
  - Drug abuse can lead to depression

Correlation $\neq$ Causality: Relationship is coincidental

- Sometimes just coincide
- Website full of examples
  `http://www.tylervigen.com/spurious-correlations`
- For example the “bald-hairy rule” in Russia: A bald (or obviously balding) state leader is succeeded by a non-bald (“hairy”) one, and vice versa:
  `https://en.wikipedia.org/wiki/Bald-hairy`

Correlation $\neq$ Causality: Relationship is coincidental

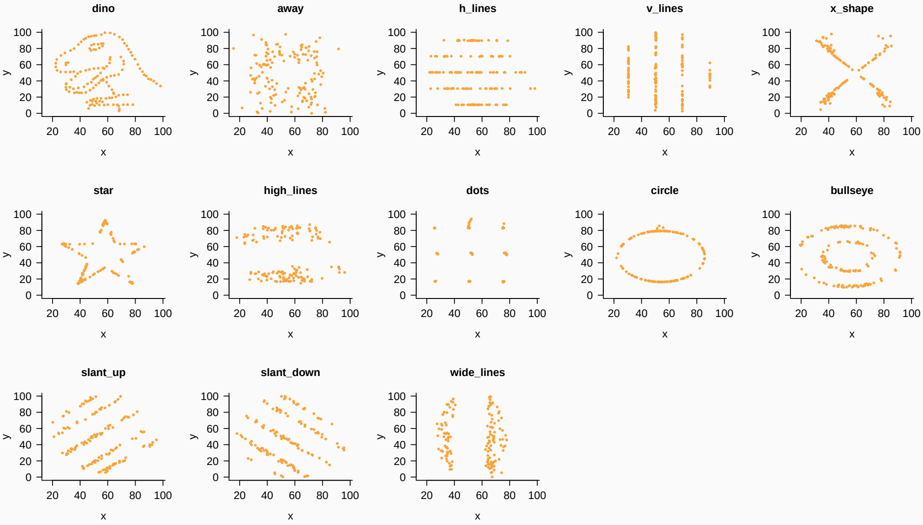


`http://www.tylervigen.com/spurious-correlations`

Beware: Coming first may not mean causality!

- *Post hoc ergo propter hoc* fallacy: “after this, therefore, because of this” fallacy
- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?

Side note: correlation is a poor indicator of how variables relate in may cases



dino

away

h_lines

v_lines

x_shape

star

high_lines

dots

circle

bullseye

slant_up

slant_down

wide_lines

No correlation does not mean no causality/dependence: “Behavioral compensation”

- Sailor sails her sailboat across a lake
- Wind blows, and she perfectly counters by turning the rudder
- Outside observer “Look at the way she’s moving that rudder back and forth but going in a straight line. That rudder is broken”
- They’re wrong but why are they wrong? There is, after all, no correlation

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

The experimental ideal

- Randomly assign T_i

The experimental ideal

- Randomly assign T_i
- T_i is independent of Y_{0i} and Y_{1i}

The experimental ideal

- Randomly assign T_i
- T_i is independent of Y_{0i} and Y_{1i}

$$\underbrace{\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)}_{\text{Observed difference}} \quad =$$

The experimental ideal

- Randomly assign T_i
- T_i is independent of Y_{0i} and Y_{1i}

$$\begin{aligned}\underbrace{\mathbb{E}(Y_i|T_i=1)-\mathbb{E}(Y_i|T_i=0)}_{\text{Observed difference}} &= \mathbb{E}(Y_{1i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=1)+\\ &\mathbb{E}(Y_{0i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=0)\end{aligned}$$

The experimental ideal

- Randomly assign T_i
- T_i is independent of Y_{0i} and Y_{1i}

$$
\begin{aligned}
\underbrace{\mathbb{E}(Y_i|T_i=1)-\mathbb{E}(Y_i|T_i=0)}_{\text{Observed difference}} &= \mathbb{E}(Y_{1i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=1)+ \\
&\quad \mathbb{E}(Y_{0i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=0) \\
&= \underbrace{\mathbb{E}(Y_{1i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=1)}_{\text{average treatment effect on the treated}}+ \\
&\quad \underbrace{\mathbb{E}(Y_{0i}|T_i=1)-\mathbb{E}(Y_{0i}|T_i=0)}_{\text{selection bias}}
\end{aligned}
$$

The experimental ideal

- Randomly assign T_i

The experimental ideal

- Randomly assign T_i
- $\mathbb{E}(Y_{0i}|T_i = 0) = \mathbb{E}(Y_{0i}|T_i = 1)$ and selection bias goes away

The experimental ideal

- Randomly assign T_i
- $\mathbb{E}(Y_{0i}|T_i = 0) = \mathbb{E}(Y_{0i}|T_i = 1)$ and selection bias goes away
- Since $\mathbb{E}(Y_{1i} - Y_{0i}|T_i = 1) = \mathbb{E}(Y_{1i} - Y_{0i})$, can drop the “on the treated” qualifier

The experimental ideal

- Randomly assign T_i

- $\mathbb{E}(Y_{0i}|T_i = 0) = \mathbb{E}(Y_{0i}|T_i = 1)$ and selection bias goes away

- Since $\mathbb{E}(Y_{1i} - Y_{0i}|T_i = 1) = \mathbb{E}(Y_{1i} - Y_{0i})$, can drop the "on the treated" qualifier

$$\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0) \quad = \quad \underbrace{\mathbb{E}(Y_{1i} - Y_{0i})}_{\text{average treatment effect (ATE)}}$$

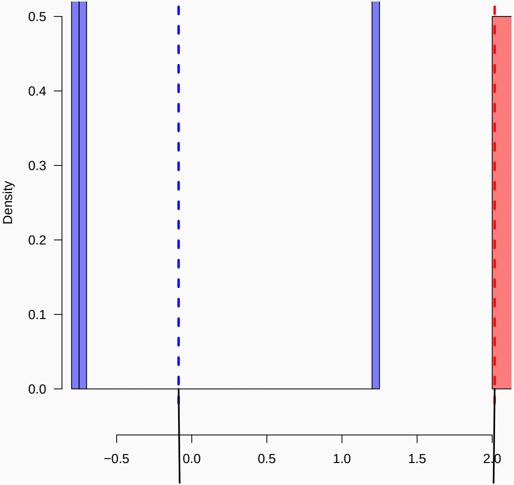
The experimental ideal

- Randomization does not eliminating individual difference (we still can't identify δ_i)
- On average, individuals in treatment/control are similar (law of large numbers)
- Need Stable Unit Treatment Value Assumption (SUTVA): Potential outcomes for any unit do not vary with the treatments assigned to other units (more later)

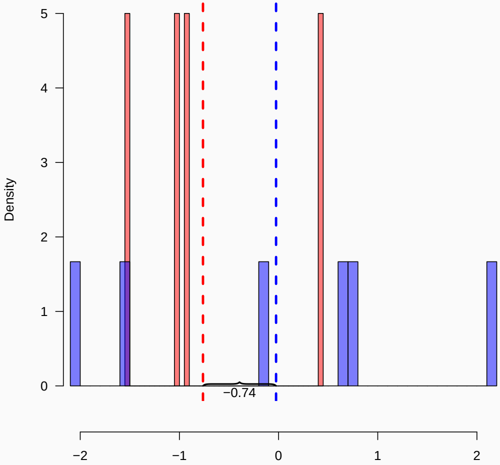
The experimental ideal: law of large numbers

```
mu0=0 #mean of untreated
s.sq=1 #variance of the outcomes
beta=0.5 #treatment effect
N=4 #Number of individuals
#Let’s create potential outcomes
Y0 <-  rnorm(n=N, mean=mu0, sd=s.sq) # control potential outcome
Y1 <- Y0 + beta # treatment potential outcome
#Lets randomly assign people to treatment
Z.sim <- rbinom(n=N, size=1, prob=.5) # Do a random assignment
Y.sim <- Y1*Z.sim + Y0*(1-Z.sim) # Reveal outcomes according to assignment
#estimate the treatment effect
mean(Y.sim[Z.sim==1])-mean(Y.sim[Z.sim==0])
hist(Y.sim[Z.sim==1],col=rgb(1,0,0,0.5),breaks=50,freq=F,
     las=1,xlab="Outcomes",xlim=range(Y.sim))
hist(Y.sim[Z.sim==0],add=T,col=rgb(0,0,1,0.5),breaks=50,freq=F)
abline(v=mean(Y.sim[Z.sim==1]),col=rgb(1,0,0,1),lwd=3,lty=2)
abline(v=mean(Y.sim[Z.sim==0]),col=rgb(0,0,1,1),lwd=3,lty=2)
```

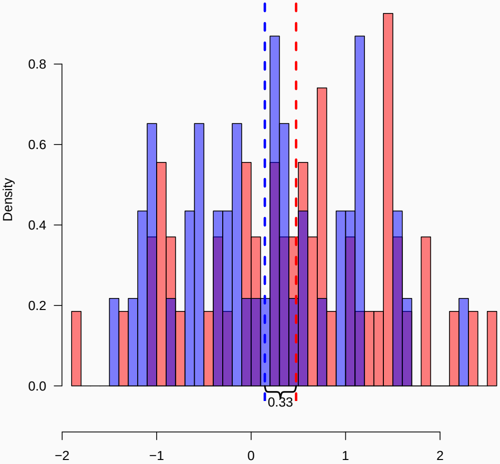
N=4



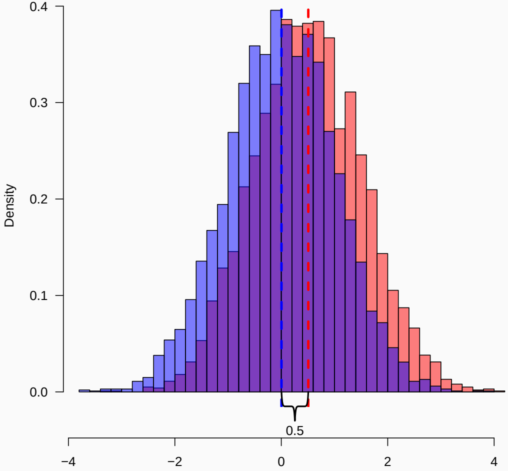
N=10



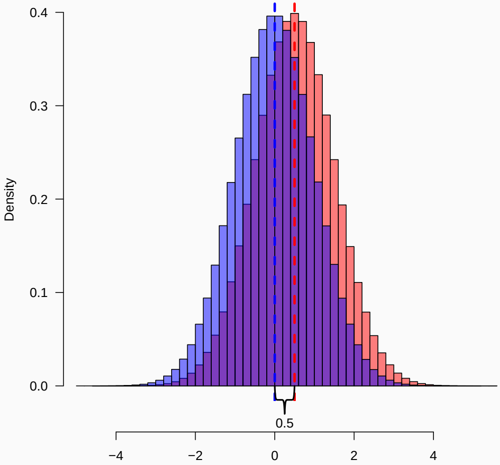
N=100



N=10,000



N=1,000,000



Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Careful with the notation

- Independence implies that **average** values for potential outcome (i.e., Y_{i1} or Y_{i0}) are the same across treatment and control

- Independence does **not** does not imply

$$E[Y_{i1}|T_i = 1] = E[Y_{i0}|T_i = 0]$$

SUTVA

- The “stable unit-treatment value assumption”
    1. **S**: *stable*
    2. **U**: across all u*nits*, or the population
    3. **TV**: t*reatment-value* (“treatment effect”, “causal effect”)
    4. **A**: a*ssumption*
- SUTVA essentially implies/assumes:
    1. homogenous dosage (treatments effect is the same for everyone)
    2. potential outcomes invariant to who else is (and how many are) treated (i.e., no externalities)
    3. partial equilibrium

SUTVA: Homogenous dose (“stable unit” part)

- Individuals are receiving the same treatment – i.e., the “dose” of the treatment to each member of the treatment group is the same
- If we are estimating the effect of hospitalization on health status, we assume everyone is getting the same dose of the hospitalization treatment.
- Easy to imagine violations if hospital quality varies across individuals
- Have to be careful what we are and are not defining as *the treatment*

SUTVA: No spillovers to other units

- Imagine treatment ($T = 1$) is vaccinating for small pox
- If i is vaccinated for small pox, then j's potential health status may be higher
- In other words, Y_j^0 may vary with T_i *regardless of* T_j
- SUTVA means that you don't have a problem like this
- If there are no externalities from treatment, then δ_i is stable for each i unit regardless of whether someone else receives the treatment too

SUTVA: Partial equilibrium only

- External validity
- Let's say we estimate a causal effect of early childhood intervention in some area
- Now adopt it for the whole country – will it have the same effect as we found?
  - Expansion may create general equilibrium effects
  - Have different effects due to economies of scale
  - The effect might be different if the population is different

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

Potential outcomes and the experimental ideal

Our goal

Potential outcomes

Correlation, causation and false counterfactuals

Experimental ideal

Randomization and selection bias

Randomization inference

The Lady Tasting Tea

Chapter II of Fisher's The Design of Experiments begins: A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

The Lady Tasting Tea

Chapter II of Fisher's The Design of Experiments begins: A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

- The lady was biologist Muriel Bristol, who worked with Fisher at the Rothamsted Experimental Station in Harpenden, UK
- H0: Fisher believes that Dr. Bristol cannot taste the difference
- A test of the hypothesis: Experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment

Randomization

- Critical assumption: if Dr. Bristol is unable to detect whether the milk was poured in first, then she will choose 4 cups at random
    - Fisher points out that the experimenter could screw this up: If all those cups made with the milk first had sugar added, while those made with the tea first had none, this might well ensure that all those made with sugar should be classed alike
    - Gerber and Green refer to this as excludability (or treatment uncorrelated with potential outcomes, but this is the whole point of randomization!)
    - To minimize the likelihood of accidentally confounding your treatment, the best approach is to constrain yourself by randomizing

The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

- Suppose Dr. Bristol correctly identified all 4 “treated” cups
    - How likely is it that this outcome could have occurred by chance?
    - There are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ possible ways to choose 4 cups
    - Only one is correct; a subject with no ability to discriminate between treated and untreated cups would have a $1/70$ chance of success
    - The p-value associated with this outcome is $1/70 \approx 0.014$

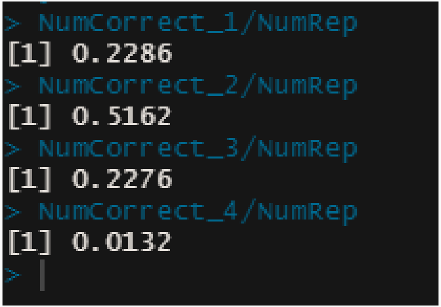
The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

- Suppose Dr. Bristol correctly identified 3 “treated” cups
  - How likely is it that this outcome could have occurred by chance?
  - There are $\binom{4}{3} \times \binom{4}{1} = 16$ possible ways to choose 3 correct cups
  - The p-value associated with this outcome is $16/70 \approx 0.22$

If you don't know the math: Simulate!

```r
NumRep=5000 #number of times we will repeat the experiment
Cups=rep(c("Milk", "Tea"), each= 4) #Make the tea cups
Cups=Cups[sample(1:8)] #Randomly order them
##Let's see how likely it is that the lady chosses X correct cups out of 4
NumCorrect_1=0
NumCorrect_2=0
NumCorrect_3=0
NumCorrect_4=0
for(r in 1:NumRep){
  #Randomly choose 4
  Selection=sample(Cups,4)
  #Count how many are correct (i.e., they are milk first)
  Correctas=sum(Selection=="Milk")
  if(Correctas==1) NumCorrect_1=NumCorrect_1+1
  if(Correctas==2)NumCorrect_2=NumCorrect_2+1
  if(Correctas==3) NumCorrect_3=NumCorrect_3+1
  if(Correctas==4) NumCorrect_4=NumCorrect_4+1
}
NumCorrect_1/NumRep
NumCorrect_2/NumRep
NumCorrect_3/NumRep
NumCorrect_4/NumRep
```



```
> NumCorrect_1/NumRep
[1] 0.2286
> NumCorrect_2/NumRep
[1] 0.5162
> NumCorrect_3/NumRep
[1] 0.2276
> NumCorrect_4/NumRep
[1] 0.0132
> 
```

The Lady Tasting Tea: a Hypothesis Test

- The only experimental result that would lead to the rejection of the hypothesis the lady was choosing at random was if she correctly identified of all 4 treated cups

- Historical note: In the actual experiment, the hypothesis that the lady was choosing at random was rejected

Randomization inference (RI)

- Elegant precursor to OLS approach to experiments by Neyman (and later by Fisher) in the 1920s
- Randomization creates a **sharp** null hypothesis ($H_o = \delta_i = 0$ for all i)
- Repeat the **exact** routine by which the original randomization was conducted many times (e.g., 5,000)
  - For each placebo treatment assignments calculate the treatment-control "statistic"
  - The distribution gives the variation in the "statistic" under the null
  - Place observed "statistic" into this distribution
  - If observed "statistic" lies in bottom or top $\frac{\alpha}{2}$% of the distribution, reject the two-sided null at α% level
- Purest form of RI: Calculate every permutation of the treatment assignment

Advantages of RI

- Allows for calculation of tight error bounds without resorting to asymptotics: **Works well in small samples**
- Makes no parametric assumptions about error distributions

6-step guide to randomization inference

1. Choose a sharp null hypothesis (e.g., no treatment effects)

2. Calculate a test statistic (S based on T and Y)

3. Then pick a randomized treatment vector $\tilde{T}_1$

4. Calculate the test statistic associated with $(\tilde{T}_1, Y)$

5. Repeat steps 3 and 4 for all possible combinations to get $\tilde{S} = \{\tilde{S}_1, \ldots, \tilde{S}_K\}$

6. Calculate exact p-value (one sided test) as $p = \frac{1}{K} \sum_{k=1}^{K} I(\tilde{S}_k \geq S)$

Pretend experiment

Pretend DBT intervention for some homeless population

Name	T	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

- For concreteness, assume a program where we pay homeless people $15 to take dialectical behavioral therapy

Step 1: Sharp null of no effect

Fisher's Sharp Null Hypothesis
$H_0 : \delta_i = Y_i^1 - Y_i^0 = 0 \ \forall i$

- Assuming no effect means any test statistic is due to chance
- Since under the Fisher sharp null $\delta_i = 0$, it means each unit's potential outcomes under both states of the world are the same
- We therefore know each unit's missing counterfactual
- The randomization we will perform will cycle through all treatment assignments under a null well treatment assignment doesn't matter because all treatment assignments are associated with a null of zero unit treatment effects
- We are looking for evidence *against* the null

Step 1: Fisher's sharp null and missing potential outcomes

- Fisher sharp null allows us to fill in the missing counterfactuals
- Under the null there's zero treatment effect at the unit level
- This guarantees zero ATE

Step 1: Fisher's sharp null and missing potential outcomes

Missing potential outcomes are no longer missing

Name	T	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Step 2: Choosing a test statistic

Test Statistic
A test statistic $S(T, Y)$ is a scalar quantity calculated from the treatment assignments T and the observed outcomes Y

- By scalar, I mean a number (vs a function) measuring some relationship between T and Y
- Ultimately there are many tests to choose from; I'll review a few later
- If you want a test statistic with high statistical power, you need large values when the null is false, and small values when the null is true (i.e., *extreme*)

Simple difference in means

- Consider the difference between treatment and control individuals

$$\delta = \frac{1}{N_T} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N_C} \sum_{i=1}^{N} (1 - T_i) Y_i$$

- Larger values of δ are evidence *against* the sharp null
- Good estimator for constant, additive treatment effects and relatively few outliers

Step 2: Calculate test statistic, $S(T, Y)$

Missing potential outcomes are no longer missing

Name	T	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

$S(T, Y) = \delta = 34/4 - 30/4 = 1$

Steps 3-5: Null randomization distribution

- Randomization steps reassign treatment assignment for every combination, calculating test statistics each time, to obtain the entire distribution of counterfactual test statistics
- The key insight of randomization inference is that under Fisher's sharp null, the treatment assignment shouldn't matter
- Ask yourself:
  - If there is no treatment effect, what must average 'placebo" test statistics equal?
  - If there is no treatment effect, what is the distribution of the 'placebo" test statistics?

Step 6: Calculate “exact” p-values

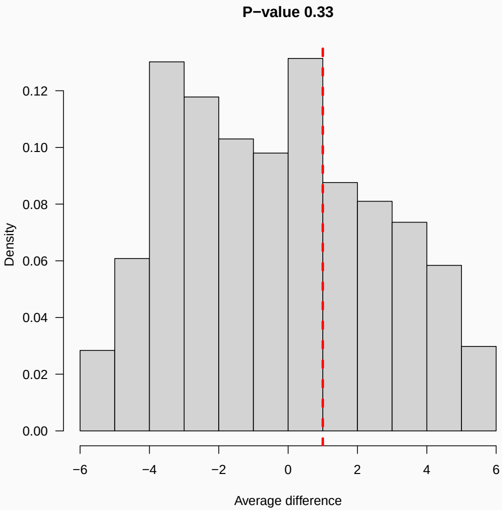
- How often would we get a test statistic as big or bigger (in absolute value) as our “real” one if sharp null was true?

- This can be calculated “easily” once we have the randomization distribution from steps 3-5

Doing randomization inference in practice

```
#Example homeless
Y=c(10,5,16,3,5,7,8,10)
Treatment=c(1,1,1,1,0,0,0,0)
RealDifference=mean(Y[Treatment==1])-mean(Y[Treatment==0])
PlaceboDifferences=NULL
NumRep=5000
for(r in 1:NumRep){
  PlaceboTreatment=sample(Treatment)
  PlaceboDifferences=c(PlaceboDifferences,
  mean(Y[PlaceboTreatment==1])-mean(Y[PlaceboTreatment==0]))
}
pvalue=mean(PlaceboDifferences>RealDifference)
hist(PlaceboDifferences,
     freq=F,las=1,xlab="Difference",
     main=paste("P-value",format(pvalue,digits=2)))
abline(v=RealDifference,col=rgb(1,0,0,1),lwd=3,lty=2)
```

Distribution of “placebo” differences, and real difference



Step 2: Other test statistics

- The simple difference in means is fine when effects are additive, and there are few outliers in the data

- What are some alternative test statistics?

Transformations

- What if there was a constant multiplicative effect: $Y_i^1 / Y_i^0 = C$?
- Difference in means will have low power to detect this alternative hypothesis
- So we transform the observed outcome using the natural log:

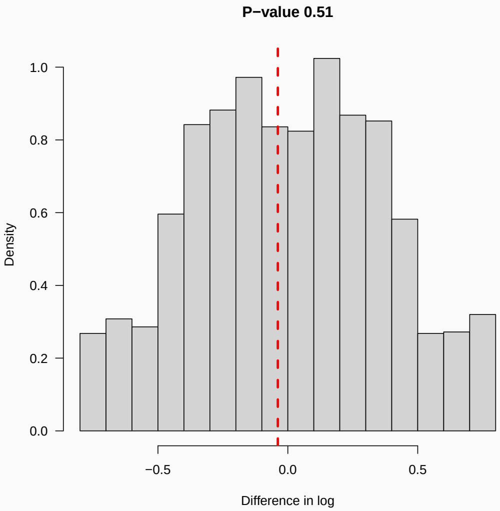
$$T_{log} = \frac{1}{N_T} \sum_{i=1}^{N} D_i ln(Y_i) - \frac{1}{N_C} \sum_{i=1}^{N} (1 - D_i) ln(Y_i)$$

- This is useful for skewed distributions of outcomes

Doing randomization inference in practice

```r
#Example homeless
Y=c(10,5,16,3,5,7,8,10)
Treatment=c(1,1,1,1,0,0,0,0)
RealDifference=mean(log(Y[Treatment==1]))-mean(log(Y[Treatment==0]))
PlaceboDifferences=NULL
NumRep=5000
for(r in 1:NumRep){
  PlaceboTreatment=sample(Treatment)
  PlaceboDifferences=c(PlaceboDifferences,
  mean(log(Y[PlaceboTreatment==1]))-mean(log(Y[PlaceboTreatment==0])))
}
pvalue=mean(PlaceboDifferences>RealDifference)
hist(PlaceboDifferences,
     freq=F,las=1,xlab="Difference",
     main=paste("P-value",format(pvalue,digits=2)))
abline(v=RealDifference,col=rgb(1,0,0,1),lwd=3,lty=2)
```

Distribution of “placebo” log differences, and real difference



Difference in medians/quantiles

- We can protect against outliers using other test statistics such as the difference in quantiles

- Difference in medians:

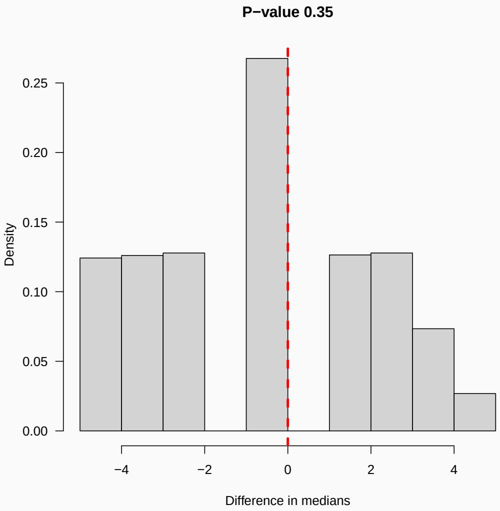
$$T_{median} = median(Y_T) - median(Y_C)$$

- We could also estimate the difference in quantiles at any point in the distribution (e.g., 25th or 75th quantile)

Doing randomization inference in practice

```r
#Example homeless
Y=c(10,5,16,3,5,7,8,10)
Treatment=c(1,1,1,1,0,0,0,0)
RealDifference=median(Y[Treatment==1])-median(Y[Treatment==0])
PlaceboDifferences=NULL
NumRep=5000
for(r in 1:NumRep){
  PlaceboTreatment=sample(Treatment)
  PlaceboDifferences=c(PlaceboDifferences,
  median(Y[PlaceboTreatment==1])-median(Y[PlaceboTreatment==0]))
}
pvalue=mean(PlaceboDifferences>RealDifference)
hist(PlaceboDifferences,
     freq=F,las=1,xlab="Difference",
     main=paste("P-value",format(pvalue,digits=2)))
abline(v=RealDifference,col=rgb(1,0,0,1),lwd=3,lty=2)
```

Distribution of “placebo” differences in medians, and real difference



Which statistic is better?

A good test statistic is the one that best fits your data

One-sided or two-sided?

- So far, we have defined all test statistics as one sides test
- We are testing against one-sided alternative

$$\begin{aligned} H_0 &: \delta_i = 0 \ \forall i \\ H_1 &: \delta_i > 0 \texttt{ for some } i \end{aligned}$$

- What about a two-sided alternative hypothesis

$$\begin{aligned} H_0 &: \delta_i = 0 \ \forall i \\ H_1 &: \delta_i \neq 0 \texttt{ for some } i \end{aligned}$$

- For these, use a test statistic that is bigger in absolute value

$$S_{diff} * = |\overline{Y}_T - \overline{Y}_C|$$