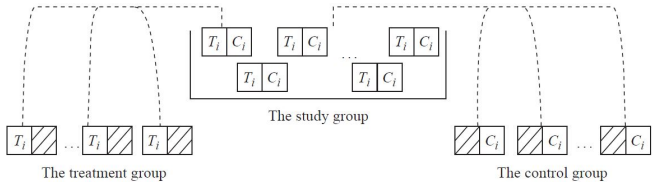
# Introduction to Empirical Economics

## *The Neyman-Rubin Causal Model*

Léo Zabrocki
leo.zabrocki@psemail.eu
https://lzabrocki.github.io/

École Normale Supérieure




The Neyman model.

# Materials of the course

https://github.com/lzabrocki/empirical_economics

*Who can summarize the previous class?*

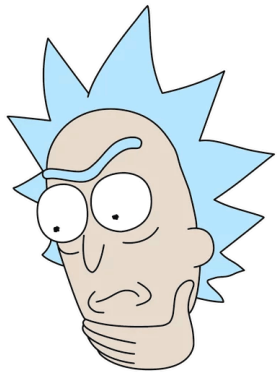
# What is causal inference?

# How to answer causal questions?

Everyday, we think about *causes* and their *effects*

Statisticians used to avoid making *causal* claims

But why?



# How to answer causal questions?

Usually, people rely on *anecdotes*

Example : efficiency of homeopathy

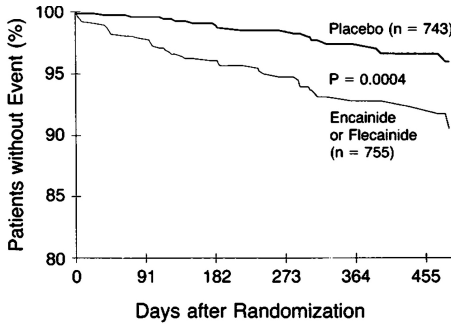
Hard to discover facts from anecdotes...

# How to answer causal questions?

Intuitions can be misleading

Treatment of Cardiac Arrhythmia is a famous example

Cardiac Arrhythmia Suppression Trial II Investigators (1992)

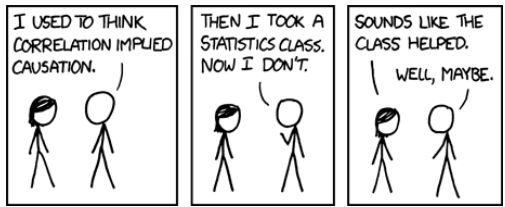


# How to answer causal questions?

Gather data and run a statistical test

Even fancy ML models can lead to biased estimates

Judea Pearl: *data are profoundly dumb*



# Randomized Experiments as a Solution

Old idea to compare similar groups

J. C. Jamison (2019) finds this idea in a letter by Francesco Petrarch (1304-1374), an Italian poet

However, no mention of a *random* allocation...

*I solemnly affirm and believe, if a hundred or a thousand men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one half followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on Nature's instincts, I have no doubt as to which half would escape.*

# Randomized Experiments as a Solution

Dutch chemist and physician Jan Baptist van Helmon (1580-1644) on bloodletting:

> *Let us take out of the Hospitals, out of the Camps, or from elsewhere, 200 or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one half of them may fall to my share, and the other to yours; I will cure them without bloodletting… we shall see how many Funerals both of us shall have.*

Again, no *random* allocation!

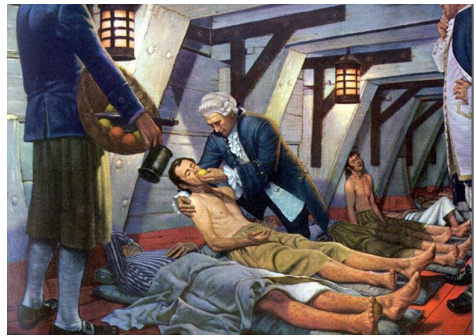
# Randomized Experiments as a Solution

James Lind, un Scottish surgeon of the Royal Navy, scurvy.

During a trip in 1747, he divded 12 sailors into 6 pairs, of which one received 2 oranges and 1 lemon per day. jour.

> *Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment of the sick in the fore-hold; and had one diet common to all.*

No *random* allocation and *control* group!



# Randomized Experiments as a Solution

Willian Waston, British physician, on smallpox (1776):

> What was the best way to inoculate patients?
>
> He step up a control group!
>
> *It was proper also to be informed of what nature unassisted, not to say undisturbed, would do for herself.*

**Table 1.** Pock Count According to the Preparatory Regimen in Children Inoculated with Smallpox.

| Pretreatment | No. of Children | Source of Inoculum* | No. of Pocks | Mean† | P Value‡ |
|---|---|---|---|---|---|
| Experiment 1 | | Early lesion | | | 0.59§ |
| Mercury plus jalap | 10 | | 25, 13, 12, 6, 5, 4, 4, 3, 0, 0 | 7.2 | |
| Senna plus rose syrup | 10 | | 30, 5, 5, 5, 4, 4, 2, 2, 0, 0 | 5.7 | |
| None | 11 | | 200, 17, 16, 16, 16, 16, 3, 2, 2, 0, 0 | 26.2 | |
| Experiment 2 | | Mature lesion | | | 0.06§ |
| Mercury | 8 | | 440, 25, 21, 21, 21, 21, 20, 7 | 72.0 | |
| Senna plus rose syrup | 8 | | 64, 26, 26, 26, 26, 26, 18, 3 | 26.9 | |
| None | 7 | | 60, 15, 15, 15, 15, 3, 2 | 17.9 | |
| Experiment 3 | | Late lesion | | | 0.09¶ |
| None | 20 | | 250, 168, 93, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 4, 4, 4, 2, 0, 0 | 51.0 | |

*The early lesion was from a patient with naturally acquired smallpox; the mature and late lesions were from inoculated patients.

†The mean was known to Watson as the "medium" and was the only calculation he could perform.

‡The P values were obtained with the Kruskal–Wallis test.

§The P value is for the comparison between either preparatory regimen and no pretreatment.

¶The P value is for the comparison of the results in the three no-pretreatment groups according to the source of the inoculum.

# Randomized Experiments as a Solution

R.A. Fisher, *The Design of Experiments* (1935) :

> Even very detailed plans of experiment cannot eliminate systematic difference between treated and control groups
>
> Solution: *random* allocation
>
> This idea took time to be well understood
>
> *At first, Fisher's use of randomisation was viewed as inconvenient, prone to error and completely unnecessary*

# The power of randomization

Randomization balances all baseline observed and *unobserved* covariates *on average*

# A Small Simulation

```
library(here) # for file paths management
library(knitr) # for generating dynamic report
library(tidyverse) # for data manipulation and visualization
library(RcppAlgos) # for computing permutations matrix
```

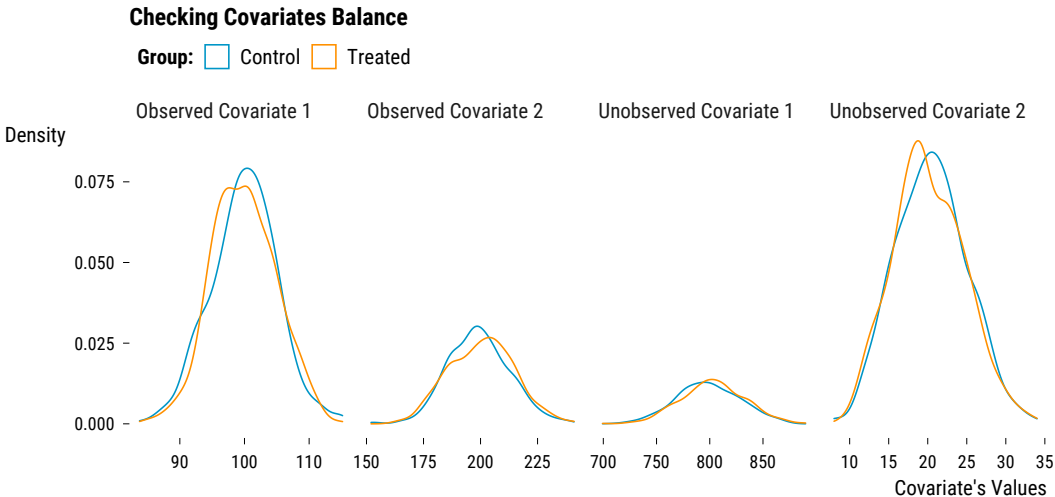
# A Small Simulation

```
# small simulation to illustrate
# the balancing property of rct

# set sample size
sample_size ← 1000

data_balance ← tibble(
  id = 1:sample_size,
  treatment = c(rep(0, sample_size/2), rep(1, sample_size/2)),
  variable_observed_1 = rnorm(n = sample_size, mean = 100, sd =5),
  variable_observed_2 = rpois(n = sample_size, lambda = 200),
  variable_unobserved_1 = rnorm(n = sample_size, mean = 800, sd = 30),
  variable_unobserved_2 = rpois(n = sample_size, lambda = 20)
)
```

# A Small Simulation

```r
graph_balance ← data_balance %>%
  # random allocation of the treatment
  mutate(treatment = sample(treatment, replace = FALSE)) %>%
  pivot_longer(
    cols = c(variable_observed_1:variable_unobserved_2),
    names_to = "covariate",
    values_to = "value"
  ) %>%
  mutate(
    covariate = case_when(
      covariate == "variable_observed_1" ~ "Observed Covariate 1",
      covariate == "variable_observed_2" ~ "Observed Covariate 2",
      covariate == "variable_unobserved_1" ~ "Unobserved Covariate 1",
      covariate == "variable_unobserved_2" ~ "Unobserved Covariate 2"
    ),
    treatment = ifelse(treatment == 1, "Treated", "Control")
  ) %>%
  ggplot(., aes(x = value, colour = treatment)) +
  geom_density() +
  scale_color_manual(values = c(my_blue, my_orange)) +
  facet_wrap( ~ covariate, scales = "free_x", nrow = 1) +
  theme_tufte() +
  labs(
    title = "Checking Covariates Balance",
    x = "Covariate's Values",
    y = "Density",
    color = "Group:"
  )
```

# A Small Simulation



# A Massive Study

J. Salk and the vaccine against polio (1955) :

- **440 000** children treated
- **210 000** received a placebo
- **1,2 million** served a control group


*Why so many children were enrolled?*

# Success of RCT in the $20^{th}$ *Century*

Some examples not in medicine:

The famous Hawthorne's experiment (1930s)

Heather Ross and the New Jersey Income Maintenance Experiment (1960s)

The Electric Company (1970s)

Was also used by many governments!

# Followed by a decline... and a come back in the 2000s

Les expériences ont été mises en arrière-plan dans les décennies qui ont suivi :

Pour revenir en force en économie au début des années 2000, notamment sur les thématiques liées au *développement* et au *marché du travail*

En 2019, Banerjee, Duflo et Kremer reçoivent le Nobel d'économie

Paradoxalement et de manière régulière, des critiques issues de la médecine se font entendre sur cette méthode

# What about COVID-19?

*[...] où j'avais invité madame Costagliola pour parler des méthodes. Nous n'étions pas d'accord, parce que sur le plan idéologique, nous ne sommes pas d'accord.*

*It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.*
*— Richard Feynman*

# The Neyman-Rubin Framework in Randomized Experiments

# The goal of statistics: *inference*

The goal of statistics is to draw conclusion from things we cannot observe using partial information.

*Inférence statistique* : learning informations on a population from a sample.

*Causal Inference* : learning what could have happened to units from what actually happened.

The goal of experiments is usually to achieve these two types of inference.

# The concept of potential outcomes

Neyman (1923) and Rubin (1974):

RCT with a binary treatment on $N$ units

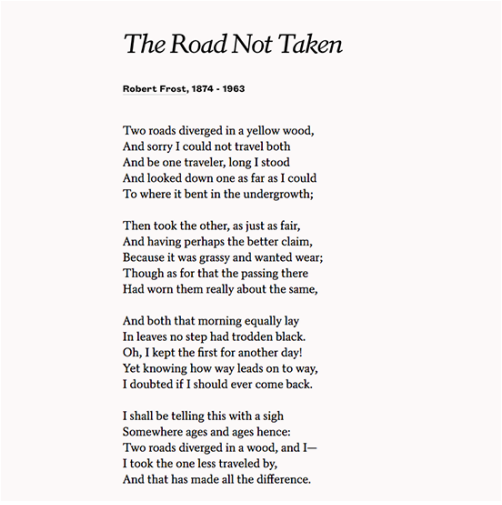
The treatment status of unit $i$ is $W_i$, which is equal to 1 when treated and 0 otherwise

Assume that each unit has two potential outcomes $Y_i(W_i = 0)$ and $Y_i(W_i = 1)$

The observed outcome of unit $i$ is therefore given by:

$$Y_i = W_i \times Y_i(1) + (1 - W_i)Y_i(0)$$

# The concept of potential outcomes



*The Road Not Taken*

**Robert Frost, 1874 - 1963**

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

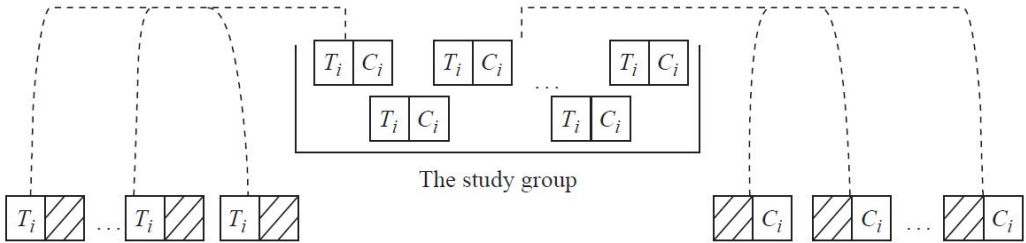
Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

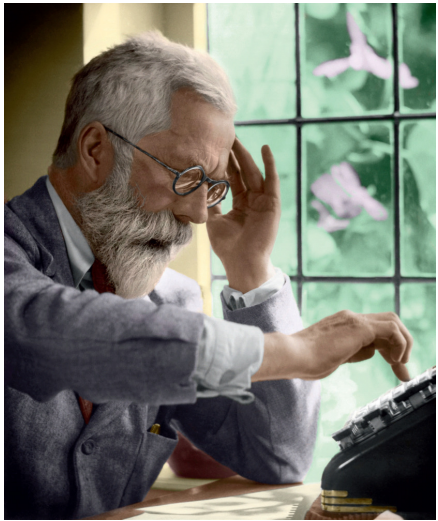
*Causal inference is inherently a missing data problem*

# The concept of potential outcomes




The Neyman model.

# A simple (historical) example



*Ronald Fisher*

*Rothamsted's Broadbalk field*

# A simple (historical) example

| Field $i$ | $W_i$ | $Y_i(0)$ | $Y_i(1)$ | $\tau_i$ | Yield (pounds) |
|---|---|---|---|---|---|
| 1 | 0 | 29.2 | ? | ? | ? |
| 2 | 0 | 11.4 | ? | ? | ? |
| 3 | 1 | ? | 26.6 | ? | 26.6 |
| 4 | 1 | ? | 23.7 | ? | 23.7 |
| 5 | 0 | 25.3 | ? | ? | 25.3 |
| 6 | 1 | ? | 28.5 | ? | 28.5 |
| 7 | 1 | ? | 14.2 | ? | 14.2 |
| 8 | 1 | ? | 17.9 | ? | 17.9 |
| 9 | 0 | 16.5 | ? | ? | ? |
| 10 | 0 | 21.1 | ? | ? | 21.1 |
| 11 | 1 | ? | 24.3 | ? | 24.3 |

*Science Table*

# Fisher's approach

Assume that sharp null hypothesis of no effect is true

$\forall i, H_0 : \tau_i = Y_i(1) - Y_i(0) = 0$

We can therefore fill the missing potential outcomes

Then, use the randomization as the *'reasoned basis for inference'*

Complete experiment with $\binom{11}{6}$=462 possible allocations

# Fisher's approach

```r
# create the data
data ← tibble(
    w = c(0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1),
    yield = c(29.2, 11.4, 26.6, 23.7, 25.3, 28.5, 14.2, 17.9, 16.5, 21.1, 24.3)
  )
```

# Fisher's approach

```
permutations_matrix ← RcppAlgos::permuteGeneral(c(0,1), 11, freq = c(5,6)) %>%
  t()
```

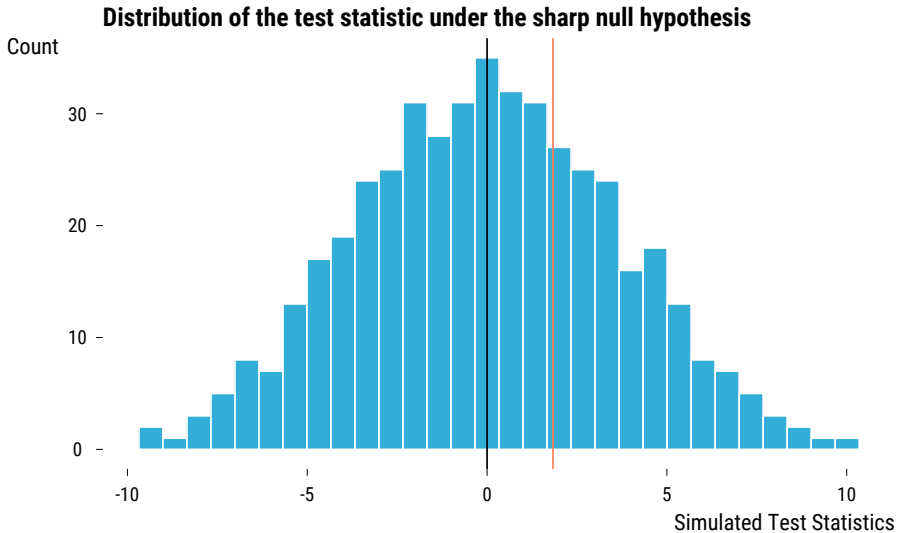
# Fisher's approach

```r
# compute observed test statistic
diff_obs ← mean(data$yield[data$w==1]) - mean(data$yield[data$w==0])

# display the value
diff_obs
```

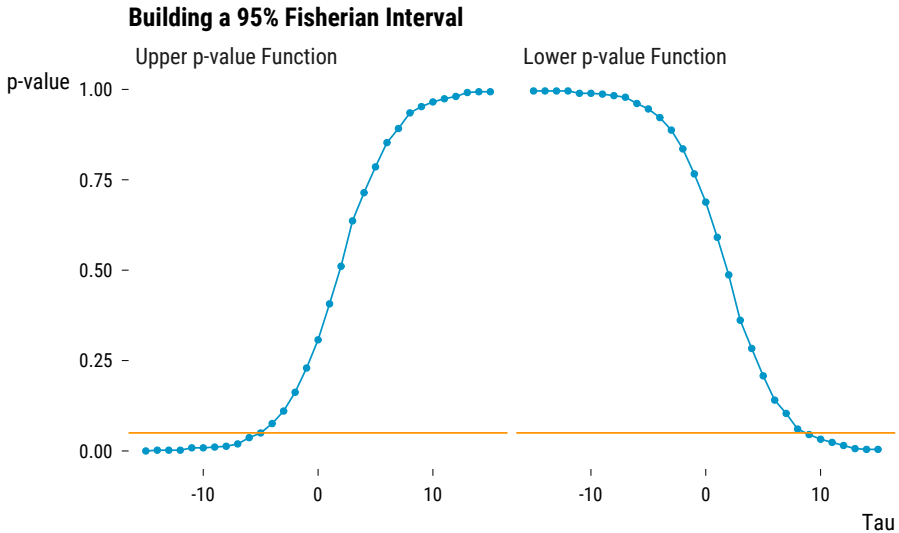
# Fisher's approach

```
# new allocation
data %>%
  mutate(w_new = sample(w, replace = FALSE)) %>%
  summarise(diff_new = mean(yield[w_new==1]) - mean(yield[w_new==0]))
```

# Fisher's approach



# Fisher's approach



# Clash at the Royal Statistical Society

**Neyman:** So long as the average yields of any treatments are identical, the question as to whether these treatments affect separate yields on single plots seems to be uninteresting

**Fisher:** It may be foolish, but that is what the z test was designed for, and the only purpose for which it has been used.

**Neyman:** I am considering problems which are important from the point of view of agriculture.

**Fisher:** It may be that the question which Dr. Neyman thinks should be answered is more important than the one I have proposed and attempted to answer. I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose.

# Neyman's approach

Neyman's was interested in estimating the *average* treatment effect:

$$ATE = E[Y_i(1) - Y_i(0)]$$

The issue is that we do not observe the unit-level treatment effects

But when we use a random allocation, we make the treatment *independent* of the potential outcomes!

$$\begin{aligned} ATE &= E[Y_i(1) - Y_i(0)] \\ &= E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0] \end{aligned}$$

We can therefore use the difference in average outcomes between the two groups as an estimator

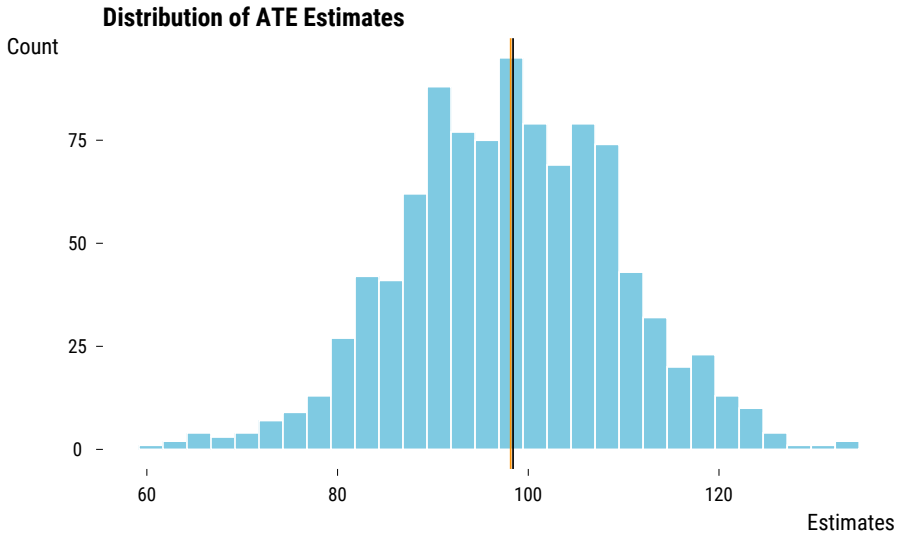
# Neyman's approach

```r
set.seed(42)

sample_size ← 1000

data_ate ← tibble(
  w = c(rep(0, sample_size/2), rep(1, sample_size/2)),
  y_0 = rnorm(n = sample_size, mean = 800, sd = 100),
  y_1 = y_0 + rnorm(n = sample_size, mean = 100, sd = 300)
)
```

# Neyman's approach

```
results_simulation ← rep(NA, 1000)

for (i in 1:length(results_simulation)) {
  data_simulation ← data_ate %>%
    mutate(w_sim = sample(w, replace = FALSE)) %>%
    mutate(y_obs = y_1 * w_sim + y_0 * (1 - w_sim))

  results_simulation[i] ←
    mean(data_simulation$y_obs[data_simulation$w_sim == 1]) - mean(data_simulation$y_obs[data_simulation$w_sim ==
                                                                                         0])
}
```

# Neyman's approach

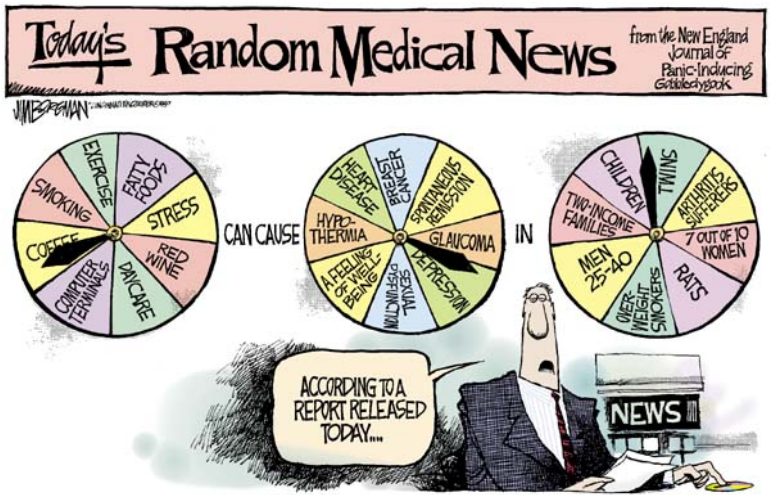


# The Neyman-Rubin Causal Model for Observational Studies

# The Issue of Selection Bias

$$E[Y_i|W_i = 1] - E[Y_i|W_i = 0] = \underbrace{E|Y_i(1)|W_i = 1] - E(Y_i(0)|W_i = 0]}_{\text{Average Treatment on the Treated}}$$
$$+ \underbrace{E[Y_i(0)|W_i = 1] - E[Y_i(0)|W_i = 0]}_{\text{Selection Bias}}$$

# The Issue of Selection Bias



# Bias in Observational Studies

**King and Zeng (2007):**

$$\begin{aligned} \text{Bias} = \ & \text{Omitted Variable Bias} + \\ & \text{Post-Treatment Bias} + \\ & \text{Interpolation Bias} + \\ & \text{Extrapolation Bias} \end{aligned}$$

# Bias in Observational Studies

**King and Zeng (2007):**

$$\text{Bias} = \text{Omitted Variable Bias} + \text{Post-Treatment Bias} + \text{Interpolation Bias} + \text{Extrapolation Bias}$$

# Bias in Observational Studies

**King and Zeng (2007):**

$$\begin{aligned} \text{Bias} = \ & \text{Omitted Variable Bias} + \\ & \text{Post-Treatment Bias} + \\ & \text{Interpolation Bias} + \\ & \text{Extrapolation Bias} \end{aligned}$$

# Bias in Observational Studies

**King and Zeng (2007):**

$$\begin{aligned} \text{Bias} = \; & \text{Omitted Variable Bias} + \\ & \text{Post-Treatment Bias} + \\ & \text{Interpolation Bias} + \\ & \text{Extrapolation Bias} \end{aligned}$$

# Bias in Observational Studies

**King and Zeng (2007):**

$$\begin{aligned} \text{Bias} = \ & \text{Omitted Variable Bias} + \\ & \text{Post-Treatment Bias} + \\ & \text{Interpolation Bias} + \\ & \text{Extrapolation Bias} \end{aligned}$$

# Bias in Observational Studies

**King and Zeng (2007):**

$$\begin{aligned} \text{Bias} = \; & \text{Omitted Variable Bias} + \\ & \text{Post-Treatment Bias} + \\ & \text{Interpolation Bias} + \\ & \text{Extrapolation Bias} \end{aligned}$$

# Cochran's Basic Advice

*How would the study be conducted if it were possible to do it by controlled experimentation?— William G. Cochran*