# Lecture 8
# Pairwise randomized experiments

# Outline

- pairwise randomized experiment
  - Fisher’s exact p-value
  - Neyman’s repeated sampling approach
  - Regression analysis
  - How to find strata / pairs?
  - R example
- Suggested reading: Imbens and Rubin Section 10.1 -10.6; Peng’s book Section 7.1-7.6

# Pairwise randomized experiment

- Procedure:
  1. Create $J = N/2$ pairs of similar units
  2. Randomize treatment assignment within each pair

- Assignment probability
  A special case of stratified randomized experiment where $N(j) = 2$ and $N_t(j) = 1$

$$P(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{X}) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} = 2^{-N/2} & \text{if } \sum_{i:B_i=j}^{N} w_i = 1 \text{ for } j = 1, \cdots, J \\ 0 & \text{otherwise} \end{cases}$$


# The Children’s television workshop experiment

[Ball, Bogatz, Rubin and Beaton, 1973.]

- The Educational Testing Service (ETS) wanted to evaluate *The Electric Company,* an American educational children's television series aimed at improving reading skills for young children

- Two sites, Yongstown, Ohio and Fresno, California where the show was not broadcast on local television, were selected to evaluate the effect of watching the show at school

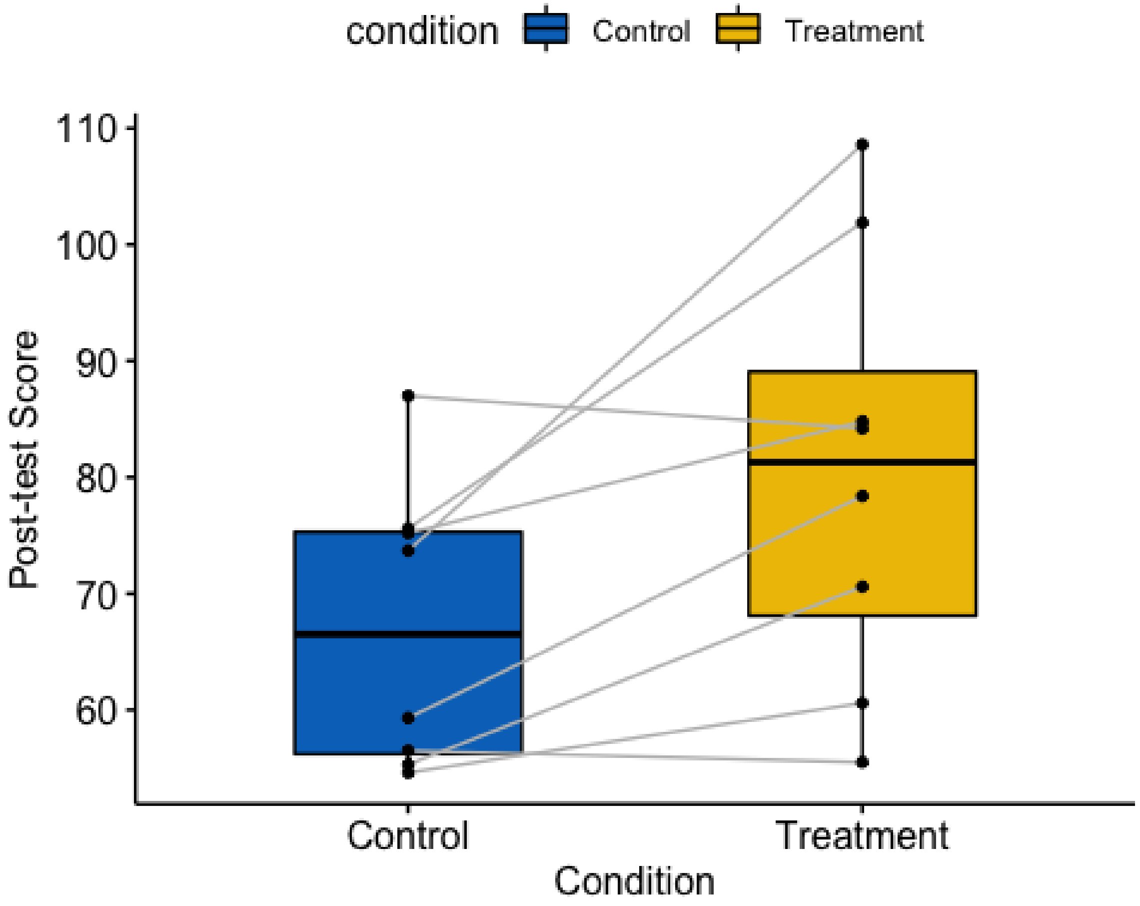
- Within each school, a pair of two classes are selected
  - One class randomly assigned to watch the show
  - Another class continue with regular reading curriculum

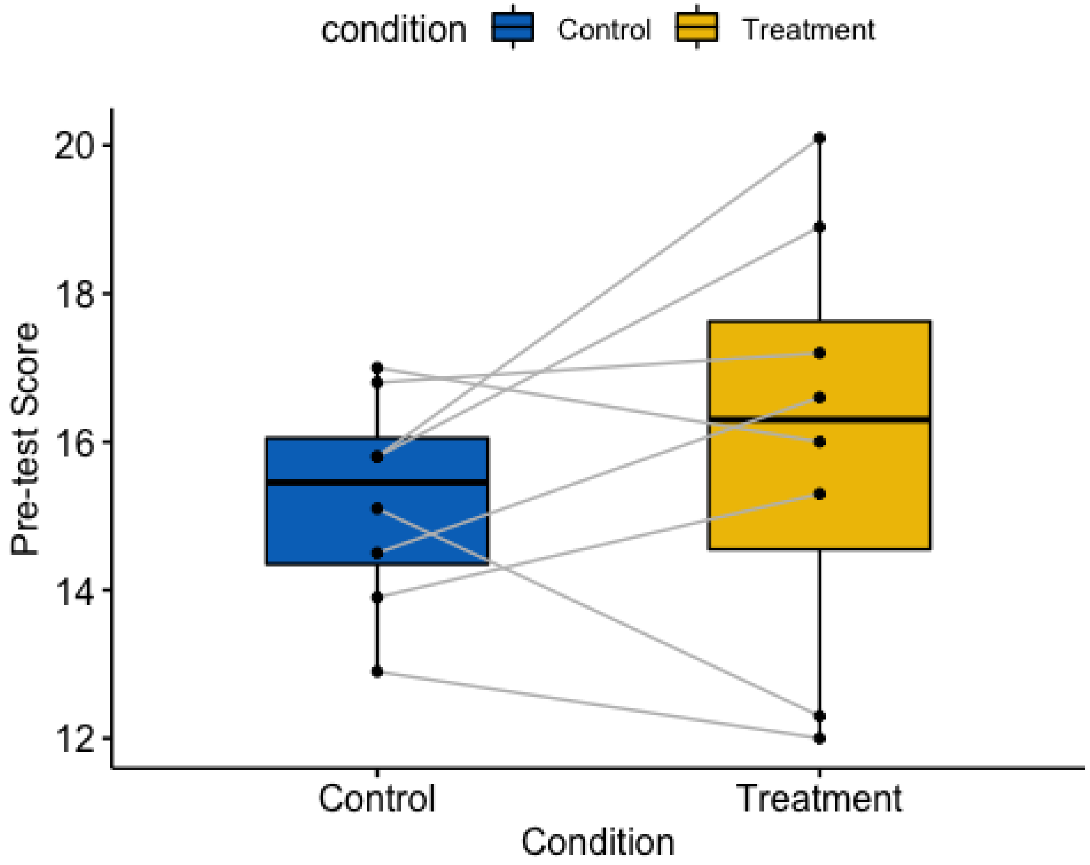
# Data from Youngstown

| Pair $G_i$ | Treatment $W_i$ | Pre-Test Score $X_i$ | Post-Test Score $Y_i^{obs}$ |
|---|---|---|---|
| 1 | 0 | 12.9 | 54.6 |
| 1 | 1 | 12.0 | 60.6 |
| 2 | 0 | 15.1 | 56.5 |
| 2 | 1 | 12.3 | 55.5 |
| 3 | 0 | 16.8 | 75.2 |
| 3 | 1 | 17.2 | 84.8 |
| 4 | 0 | 15.8 | 75.6 |
| 4 | 1 | 18.9 | 101.9 |
| 5 | 0 | 13.9 | 55.3 |
| 5 | 1 | 15.3 | 70.6 |
| 6 | 0 | 14.5 | 59.3 |
| 6 | 1 | 16.6 | 78.4 |
| 7 | 0 | 17.0 | 87.0 |
| 7 | 1 | 16.0 | 84.2 |
| 8 | 0 | 15.8 | 73.7 |
| 8 | 1 | 20.1 | 108.6 |

- Two first-grade classes from each of eight schools participate in the experiment
- ETS performed reading ability tests to the kids both before the program started and after it finished.

# Data from Youngstown



# Some notations

| Pair | Unit A | | | | | Unit B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $Y_{i,A}(0)$ | $Y_{i,A}(1)$ | $W_{i,A}$ | $Y_{i,A}^{\text{obs}}$ | $X_{i,A}$ | $Y_{i,B}(0)$ | $Y_{i,B}(1)$ | $W_{i,B}$ | $Y_{i,B}^{\text{obs}}$ | $X_{i,B}$ |
| 1 | 54.6 | ? | 0 | 54.6 | 12.9 | ? | 60.6 | 1 | 60.6 | 12.0 |
| 2 | 56.5 | ? | 0 | 56.5 | 15.1 | ? | 55.5 | 1 | 55.5 | 13.9 |
| 3 | 75.2 | ? | 0 | 75.2 | 16.8 | ? | 84.8 | 1 | 84.8 | 17.2 |
| 4 | 76.6 | ? | 0 | 75.6 | 15.8 | ? | 101.9 | 1 | 101.9 | 18.9 |
| 5 | 55.3 | ? | 0 | 55.3 | 13.9 | ? | 70.6 | 1 | 70.6 | 15.3 |
| 6 | 59.3 | ? | 0 | 59.3 | 14.5 | ? | 78.4 | 1 | 78.4 | 16.6 |
| 7 | 87.0 | ? | 0 | 87.0 | 17.0 | ? | 84.2 | 1 | 84.2 | 16.0 |
| 8 | 73.7 | ? | 0 | 73.7 | 15.8 | ? | 108.6 | 1 | 108.6 | 20.1 |

- Average treatment effect within pair $j$

$$\tau^{\text{pair}}(j) = \frac{1}{2}\sum_{i:G_i=j}\big(Y_i(1) - Y_i(0)\big) = \frac{1}{2}\Big(\big(Y_{j,A}(1) - Y_{j,A}(0)\big) + \big(Y_{j,B}(1) - Y_{j,B}(0)\big)\Big).$$

- Observed outcomes for both treatment and control groups

$$Y_{j,c}^{\text{obs}} = \begin{cases} Y_{j,1}(0) & \text{if } W_{j1} = 0 \\ Y_{j,2}(0) & \text{if } W_{j2} = 0 \end{cases} \quad \text{and} Y_{j,t}^{\text{obs}} = \begin{cases} Y_{j,1}(1) & \text{if } W_{j1} = 1 \\ Y_{j,2}(1) & \text{if } W_{j2} = 1 \end{cases}$$

# Fisher's exact p-value

- We still focus on the **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, N$
- Choice of test statistics:
  - Average group mean differences across pairs

$$T^{\text{dif}} = \left| \frac{1}{J} \sum_{j=1}^{J} \left( Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} \right) \right| = \left| \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} \right|$$

    As each pair has exactly one treatment and one control
    - We don't need to consider different weights
    - No worry of Simpson's paradox

  - Rank statistics
    - Use population ranks: $T = |\overline{\text{rank}}(Y_t^{\text{obs}}) - \overline{\text{rank}}(Y_c^{\text{obs}})|$
    - Use within-pair ranks

$$T^{\text{rank,pair}} = \left| \frac{2}{N} \sum_{j=1}^{N/2} \left( \mathbf{1}_{Y_{j,1}^{\text{obs}} > Y_{j,0}^{\text{obs}}} - \mathbf{1}_{Y_{j,1}^{\text{obs}} < Y_{j,0}^{\text{obs}}} \right) \right|$$

# Application to the television workshop data

- Fisher's exact p-values
  - Mean differences: $T = 13.4$, $\text{pvalue} = 0.031$
  - Rank mean differences: $T = 3.75$, $\text{pvalue} = 0.031$
  - Within-pair rank differences: $T = 0.5$, $\text{pvalue} = 0.29$

- Rank v.s. within-pair rank
  - Both can reduce the sensitivity to outliers
  - Using within-pair ranks can have more power when there is substantial variation in the level of the outcomes between pairs
  - Otherwise, using within-pair ranks loses power as it treats small within-pair differences (which may be due to random noises) equally with large within-pair differences

  - Using within-pair ranks is more appropriate for large, heterogenous population

# Neyman's repeated sampling approach

- Target: PATE or SATE $\tau = \sum_j \frac{N(j)}{N} \tau(j)$ where $\tau(j)$ is the PATE or SATE for strata $j$
- Point estimate:

$$\hat{\tau}^{\text{pair}}(j) = Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} \qquad \hat{\tau}^{\text{dif}} = \frac{1}{N/2} \sum_{j=1}^{N/2} \hat{\tau}^{\text{pair}}(j) = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$$

- $\mathbb{E}\left(\hat{\tau}^{\text{dif}}\right) = \tau$

$$\mathbb{E}\left(Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}}\right) = \mathbb{E}\left( W_{j1} Y_{j,1}(1) + W_{j2} Y_{j,2}(1) - \left(1 - W_{j1}\right) Y_{j,1}(0) - \left(1 - W_{j2}\right) Y_{j,2}(0) \right) = \tau^{\text{pair}}(j)$$

- We can not estimate the within-pairs variances as there are only two units per pair
- Use the following empirical estimate of the uncertainty (paired t-test)

$$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{4}{N \cdot (N-2)} \cdot \sum_{j=1}^{N/2} \left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2$$

- Above estimate is conservative
  - $\hat{\tau}^{\text{pair}}(j)$ has mean $\tau^{\text{pair}}(j)$ instead of $\tau$

$$\mathbb{E}\left[\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)\right] = \mathbb{V}_W(\hat{\tau}^{\text{dif}}) + \frac{4}{N \cdot (N-2)} \cdot \sum_{j=1}^{N/2} \left(\tau^{\text{pair}}(j) - \tau\right)^2$$

# Application to the television workshop data

- Est. = 13.4, sd. = 4.6, 95% CI: $[4.3, 22.5]$
- As we have 8 pairs, Gaussian approximation is inaccurate and it's better to compare with a t-distribution with df = 7
- 95% CI comparing with t-distribution: $[2.5, 24.3]$
- If we treat the data as from completely randomized experiment, then sd. = 7.8

| Pair | Outcome for Control Unit | Outcome for Treated Unit | Difference |
|---|---|---|---|
| 1 | 54.6 | 60.6 | 6.0 |
| 2 | 56.5 | 55.5 | −1.0 |
| 3 | 75.2 | 84.8 | 9.6 |
| 4 | 75.6 | 101.9 | 26.3 |
| 5 | 55.3 | 70.6 | 15.3 |
| 6 | 59.3 | 78.4 | 19.1 |
| 7 | 87.0 | 84.2 | −2.8 |
| 8 | 73.7 | 108.6 | 34.9 |
| Mean | 67.2 | 80.6 | 13.4 |
| (S.D.) | (12.2) | (18.6) | (13.1) |

# Linear regression

- We can not run separate linear regressions within each pair, as there are only 2 units per pair

How to build a reasonable regression framework?

- For each pair $j$, $Y_{j,k}(w) = Y_{j,k}(0) + \tau_{j,k}$ for $k = 1$ or 2
- We assume that

$$\mathbb{E}(Y_{j,k}(0)|\boldsymbol{X}) = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{X}_{j,k}, \qquad \mathbb{E}(\tau_{j,k}|\boldsymbol{X}) = \tau + \boldsymbol{\gamma}^T(\boldsymbol{X}_{j,k} - \overline{\boldsymbol{X}})$$

- Then

$$\mathbb{E}(Y_{j,k}(w)|\,\boldsymbol{X}_{jk}) = \alpha_j + \tau w + \boldsymbol{\beta}^T \boldsymbol{X}_{j,k} + w\boldsymbol{\gamma}^T(\boldsymbol{X}_{j,k} - \overline{\boldsymbol{X}})$$

- Unconfoundedness property (also implicitly condition on pair indicators):

$$(\boldsymbol{Y}(0), \boldsymbol{Y}(1)) \perp \boldsymbol{W} \mid \boldsymbol{X}$$

- Then we have

$$\mathbb{E}(Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}}|\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y_{j,t}(1) - Y_{j,c}(0)|\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{X} = \boldsymbol{x})$$
$$= \mathbb{E}(Y_{j,t}(1) - Y_{j,c}(0)|\boldsymbol{X} = \boldsymbol{x})$$

where $Y_{j,t}^{\text{obs}}$ and $Y_{j,c}^{\text{obs}}$ are observed responses for the treated and control unit in the $j$th pair

# Linear regression

- We finally have the regression model:

$$\mathbb{E}\left(Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} | \boldsymbol{W} = \boldsymbol{w}, \boldsymbol{X} = \boldsymbol{x}\right) = \mathbb{E}\left(Y_{j,t}(1) - Y_{j,c}(0) | \boldsymbol{X} = \boldsymbol{x}\right)$$
$$= \tau + \boldsymbol{\gamma}^T\left( \boldsymbol{X}_{j,t} - \overline{\boldsymbol{X}}\right) + \boldsymbol{\beta}^T\left( \boldsymbol{X}_{j,t} - \boldsymbol{X}_{j,c}\right)$$
$$= \tau + \boldsymbol{\gamma}^T\left(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}}\right) + \left(\boldsymbol{\beta} + \frac{\boldsymbol{\gamma}}{2}\right)^T \left( \boldsymbol{X}_{j,t} - \boldsymbol{X}_{j,c}\right)$$

- $\tau$ is still the PATE
- We still implicitly condition on the pair indicators variables

- If $\boldsymbol{\gamma} = \boldsymbol{0}$, then $\mathbb{E}\left(Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} | \boldsymbol{W} = \boldsymbol{w}, \boldsymbol{X} = \boldsymbol{x}\right) = \tau + \boldsymbol{\beta}^T\left( \boldsymbol{X}_{j,t} - \boldsymbol{X}_{j,c}\right)$ we only need to include the covariates differences in the linear regression model

- We can assume homoscedastic errors in the linear regression even if $\mathbb{V}(Y_i(0)) \neq \mathbb{V}(Y_i(1))$
    - We assume the pairs are i.i.d.

# How to perform stratification / pairing

- Implementation based on convenience

- Univariate blocking: discrete or discretized variable
- Multivariate blocking: Mahalanobis distance

$$D(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^\top \widehat{\mathbb{V}(\mathbf{X})}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

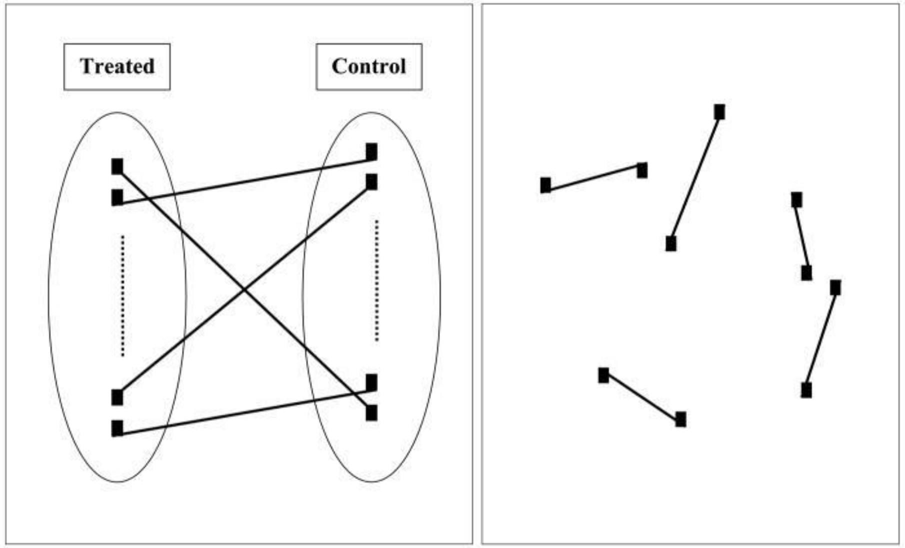
Greedy algorithms

- Matching: pair two units with the shortest distance, set them aside, and repeat
- Blocking: randomly choose one unit and choose $N_j$ units with the shortest distances, set them aside, and repeat

But the resulting matches may not be optimal

# Optimal matching

- $D$: $N \times N$ matrix of pairwise distance or a cost matrix
- Optimal matching
  - Binary $N \times N$ matching matrix: $M$ with $M_{ij} \in \{0,1\}$
  - Optimization problem:
    $\min_M \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} D_{ij}$ subject to $\sum_{i=1}^{N} M_{ij} = 1$ for all $j$
    where we set $D_{ii} = \infty$ for all $i$
  - $M$ also need to be symmetric




Bipartite Matching Nonbipartite Matching

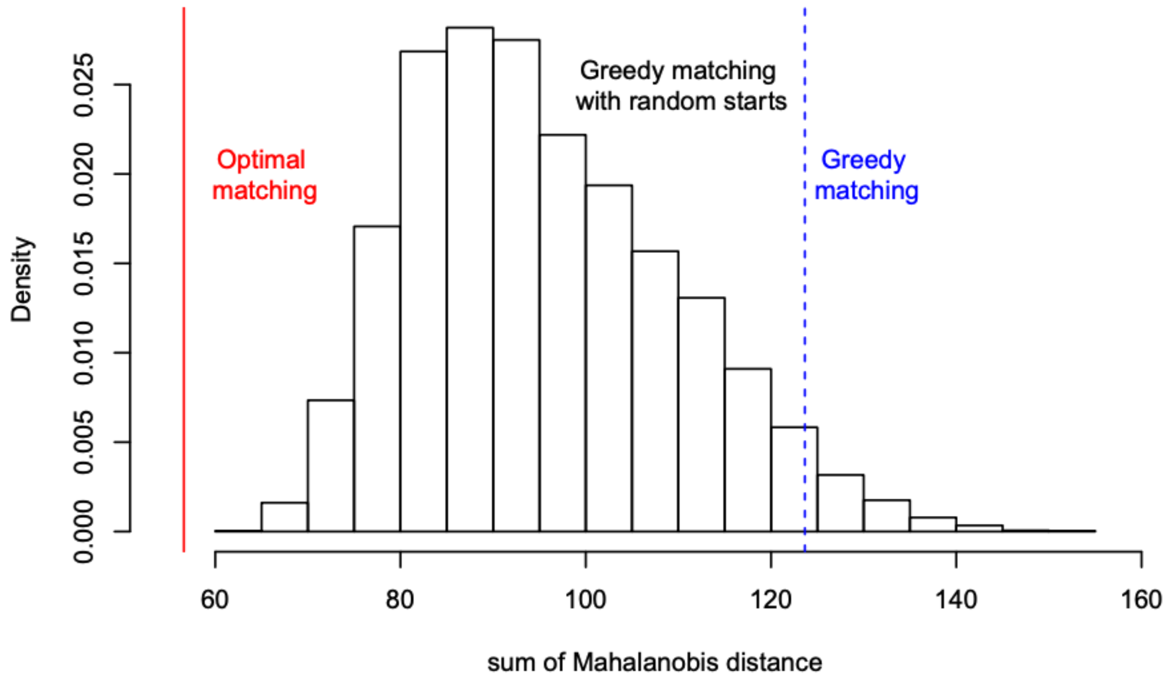
- Nonbipartite matching

- Computational cost $O(n^3)$
- Derigs' algorithm: implemented in the R package nbpMatching
  https://cran.r-project.org/web/packages/nbpMatching/

# Example: evaluation of health insurance policy

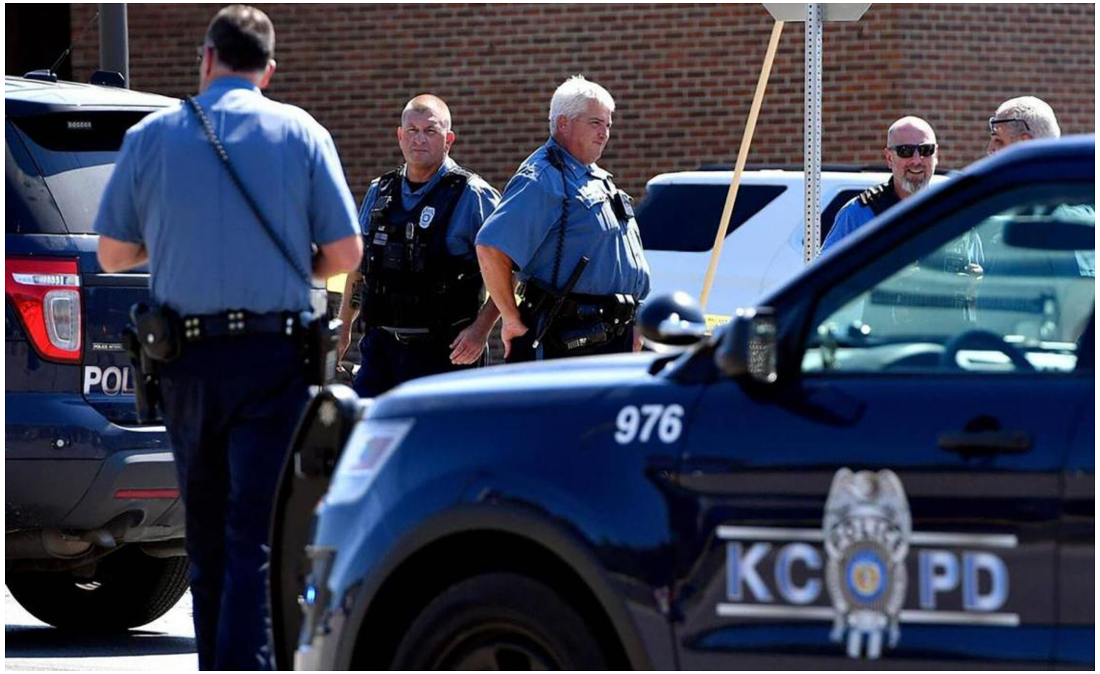
[Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. *The lancet, 2009*.]

- Seguro Popular, a programme aimed to deliver health insurance, regular and preventive medical care, medicines, and health facilities to 50 million uninsured Mexicans
- Units: health clusters = predefined health facility catchment areas
- 4 pre-treatment cluster-average covariates: age, education, household size, household assets
- 100 clusters, 50 pairs



# Case study: Kansas City Preventive Patrol Experiment

- A landmark experiment carried out between October 1, 1972, through September 30, 1973




## Goal:

- Test for two fundamental hypotheses:
  1. **Visible Police Presence Deters Crime**: potential offenders would be less likely to commit crimes if they saw police patrols.
  2. **Police Presence Reduces Public Fear**: seeing police patrols would make the community feel safer.

## Preventive patrol

police actively patrol an area in an attempt to prevent crime from occurring

# Case study: Kansas City Preventive Patrol Experiment

**Table 16:**
**PATROL IS THE MOST IMPORTANT FUNCTION IN THE POLICE DEPARTMENT**

Total Responding = 178     0 = 1.94     S.D. = 1.05

| Response of South Patrol Division Police Officers | |
|---|---|
| Strongly agree | 42.2% |
| Moderately agree | 32.8% |
| Slightly agree | 1.7% |
| Slightly disagree | 5.0% |
| Moderately disagree | 0.6% |
| Strongly disagree | 1.1% |
| No response | 16.7% |

Survey from the police

# Experimental design

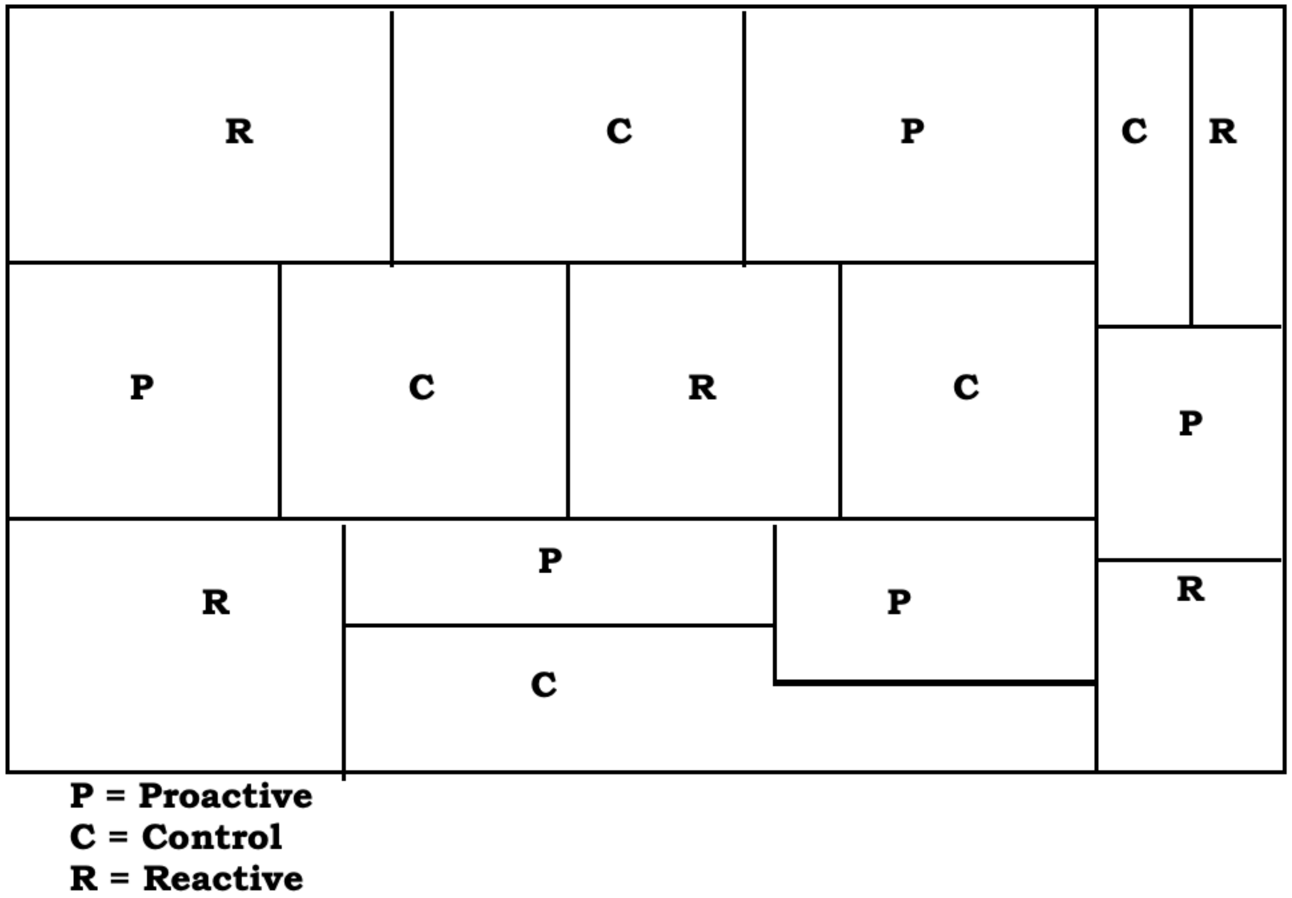
- Among South Patrol Division's 24-beat area, nine beats were eliminated as unrepresentative of the city's socioeconomic composition.

- The remaining 15 beats are computationally matched into 5 groups, 3 beats for each group

- Randomization within each group: randomly select one beat for each treatment level
  - Reactive Patrol(R): Police cars were removed from these beats. Officers only responded to calls for service.
  - Standard Patrol (C): These beats acted as the control group, with policing continuing as usual.
  - Proactive Patrol (P): Police patrols were significantly increased in these beats.

- It was agreed that if a noticeable increase in crime occurred within a reactive beat, the experiment would be suspended.
- Additional training to the police that encourage them to adhere to the treatment assignment

# Experimental design and outcome




Outcome measured

- Crime rates
- Response times
- Community attitudes toward the police

- Data are collected from community surveys, interviews, recorded observations and departmental data

# Analysis result

- Performed two-sample t-test
- no significant differences in the level of crime, citizens' attitudes toward police services, citizens' fear of crime, police response time, or citizens' satisfaction with police response time.
- Summary report available at: https://www.policinginstitute.org/wp-content/uploads/2015/07/Kelling-et-al.-1974-THE-KANSAS-CITY-PREVENTIVE-PATROL-EXPERIMENT.pdf

**Table 2: DEPARTMENTAL REPORTED CRIME**

| Crime Type | Overall P | R,C | R,P | C,P |
|---|---|---|---|---|
| Robbery - Inside | | R=C | R=P | C=P |
| Robbery - Outside | | R=C | R=P | C=P |
| Common Assault | | R=C | R=P | C=P |
| Aggravated Assault | | R=C | R=P | C=P |
| Larceny - Purse Snatch | | R=C | R=P | C=P |
| Rape | | R=C | R=P | C=P |
| Other Sex Crimes | $.01<p<.025$ | R>C | R=P | C=P |
| Homicide | | R=C | R=P | C=P |
| Residence Burglary | | R=C | R=P | C=P |
| Non-Residence Burglary | | R=C | R=P | C=P |
| Auto Theft | | R=C | R=P | C=P |
| Vandalism | | R=C | R=P | C=P |
| Larceny - Auto Accessory | | R=C | R=P | C=P |
| Larceny - Theft from Auto | | R=C | R=P | C=P |
| Larceny - Bicycle | | R=C | R=P | C=P |
| Larceny - Shoplift | | R=C | R=P | C=P |
| Larceny - Theft from Bldg. | | R=C | R=P | C=P |

# Comments on the analysis result

What can be the potential drawbacks of the experimental design and analysis?

- Data analyzed by two-sample testing, not as from paired randomized experiment, so statistical test can be conservative
- Sample size is small
- Short term effect may be small

- Non-compliance → Police presence are kept monitored during the experiment
  - However, the study did not collect data on the amount of preventive patrol in each condition (Weisburd et. al. 2023)
- Spill-over effect → Assessed by evaluating correlation between nearby beats to indicate no spill-over effect
- The randomization is questioned (Weisburd et. al. 2023): four R beats are on the corner of the region