

# Causal Inference Methods and Case Studies

STAT24630

Jingshu Wang

# Lecture 4

Topic: Classical randomized experiments

- A case study using Fisher's sharp null and exact p-values

# Case study: the California alphabet lottery

[Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American statistical association*, 2006]

## Problem background

- In the 2000 U.S. national election, George W. Bush became President by winning 537 more votes than Al Gore in Florida.
- This unusually close election result served as a reminder that the manner in which elections are administered can change outcomes.
- This paper studied the causal effect of the page placement of candidates in the 2003 California recall election
- dataset was collected by *The New York Times* in 2003 (not publicly available)

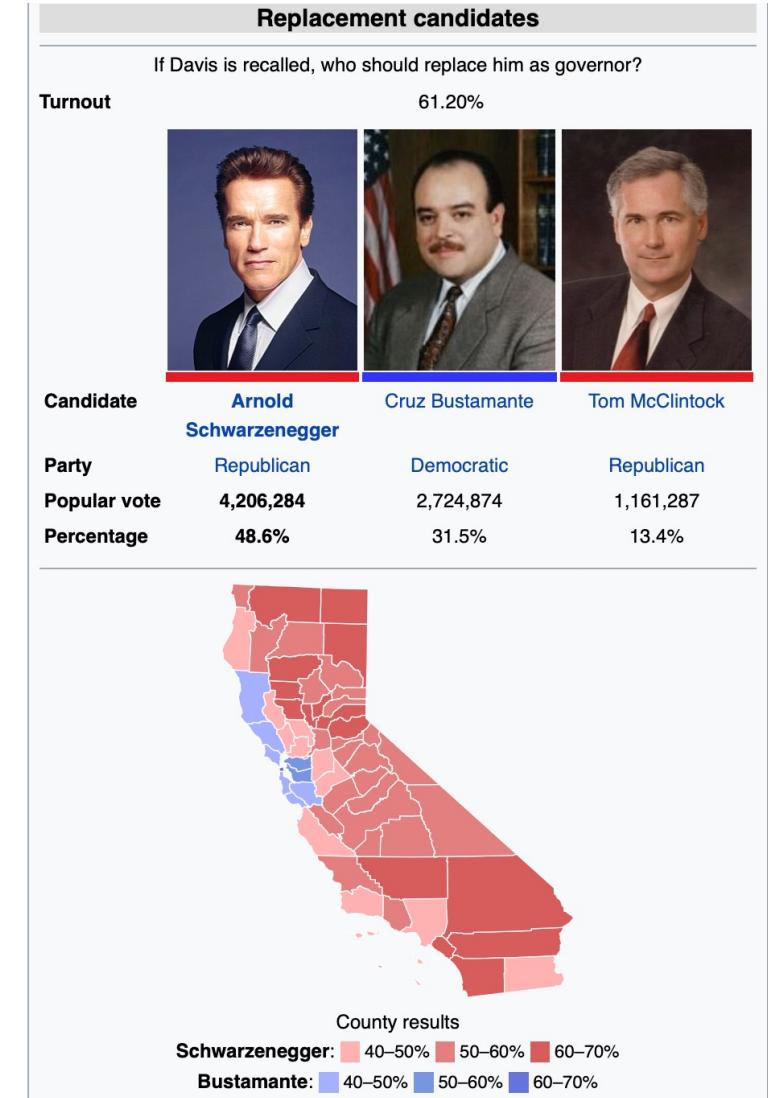
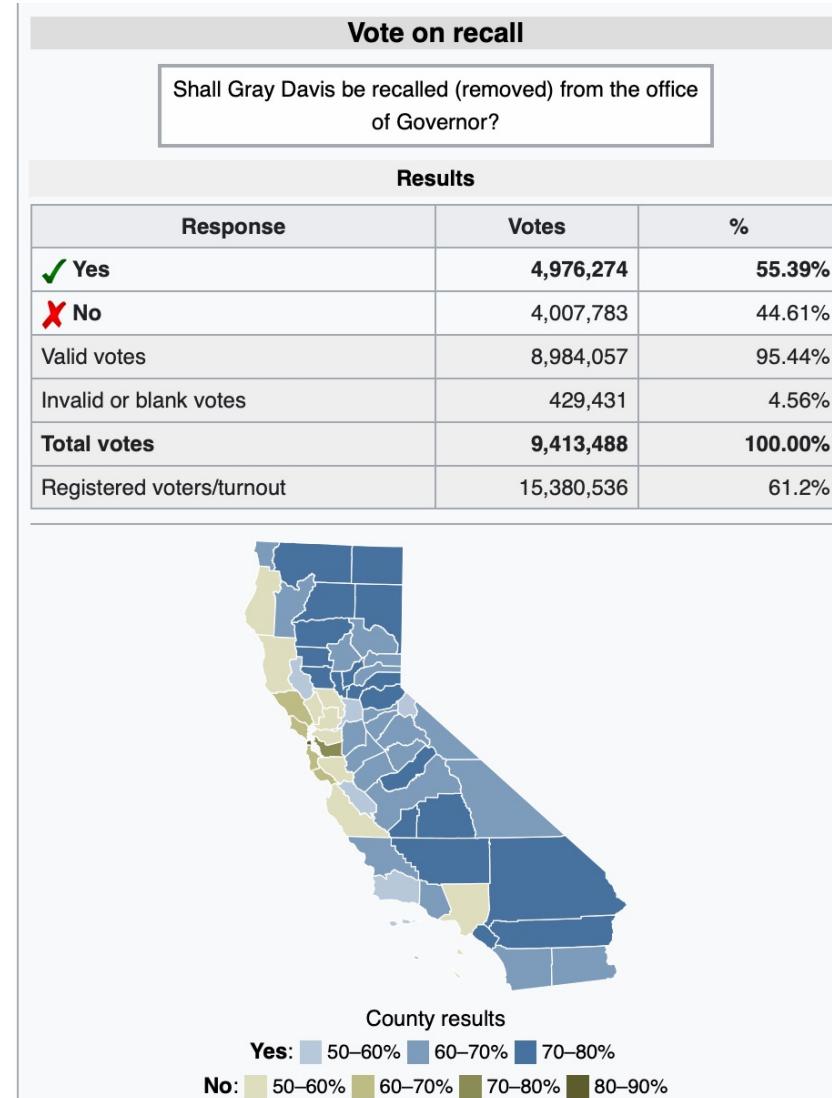
# Case study: the California alphabet lottery

[Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American statistical association*, 2006]

## Problem background

- Recall results

[https://en.wikipedia.org/wiki/2003\\_California\\_gubernatorial\\_recall\\_election](https://en.wikipedia.org/wiki/2003_California_gubernatorial_recall_election)



# The randomization-rotation procedure

- Since 1975, California law has mandated that the Secretary of State draw a random alphabet for each election to determine the order of candidates for the first assembly district [California Election Code § 13112 (2003)].
- California law further requires that the candidate order be systematically rotated throughout the remaining assembly districts.
- [The procedure](#)
  1. Randomize alphabet
  2. Sort candidates by randomized alphabet
  3. Rotate the candidate order from the first district

For the 2003 recall election, the actual randomized alphabet was

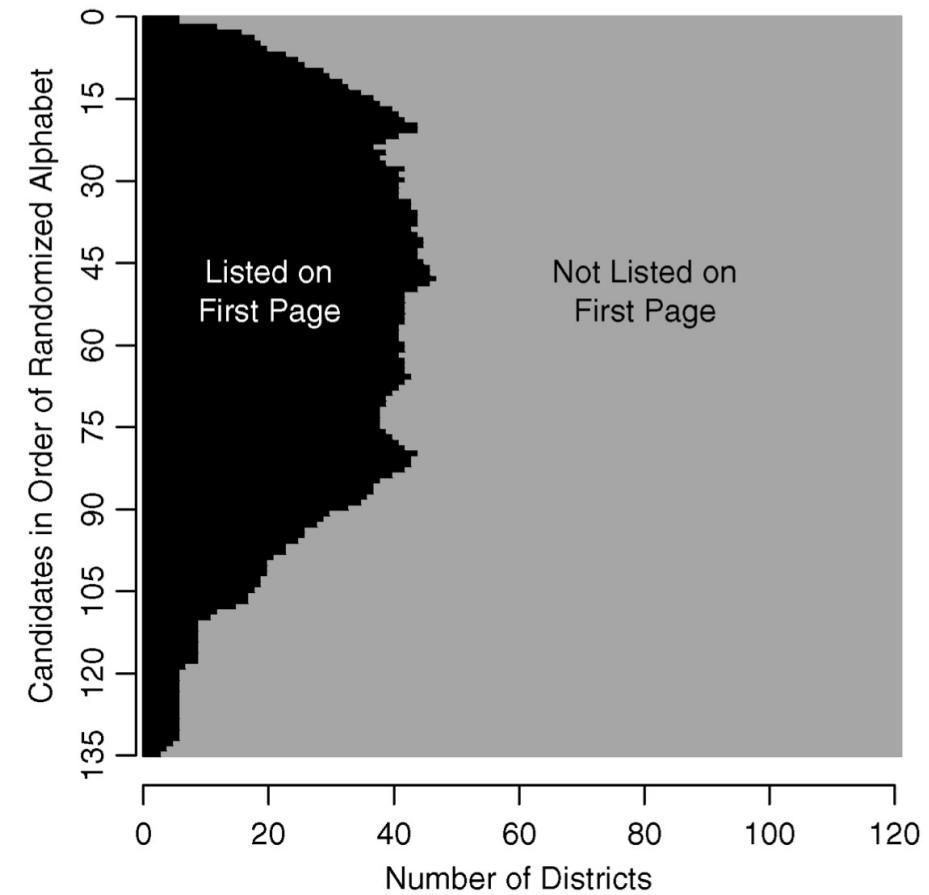
R W Q O J M V A H B S G Z X N T C I E K U P D Y F L

- The ballot order in the first assembly district was determined, starting from Robinson, Roscoe, Ramirez, and so on and proceeding to Lewis and Leonard.
- This candidate order was then rotated throughout the remaining assembly districts.

# The randomization-rotation procedure

## Challenges analyzing data with the randomization procedure

- Randomization is not done on each candidate
- The alphabets are randomized, but the 80 assembly districts order are not randomized
- an unprecedented total of 135 candidates, from Hollywood actor Arnold Schwarzenegger to child television star Gary Coleman
- Each of the 58 counties uses a different ballot format with varying numbers of pages, leading to 121 county-district combinations of ballot formats
- interactions across candidates



No complete randomization of page placement across candidates nor across districts

# Set up the analysis framework

- Analyze the causal effect of page placement for each of the 135 candidates separately
- Each of 121 county-district combination is a **unit**:  $Y_i(0)$  and  $Y_i(1)$  for a district  $i$  and a particular candidate
- Treatment:  $T_i = 1$  if candidate is placed on the first page,  $T_i = 0$  otherwise
- Sharp null for a particular candidate:  $H_0: Y_i(0) \equiv Y_i(1)$  for all  $i = 1, \dots, 121$
- Test statistics:
  - Sample average treatment effect  $W^D(\mathbf{T}) = \frac{\sum_{i=1}^{121} T_i y_i}{N_1} - \frac{\sum_{i=1}^{121} (1 - T_i) y_i}{N_0}$
  - Covariate-adjusted test statistics

$$W^L(\mathbf{T}) = (\mathbf{T}^\top \mathbf{M} \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{M} \mathbf{y}, \quad (4)$$

where  $\mathbf{y} = (y_1, y_2, \dots, y_{121})$ ,  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ , and  $\mathbf{X}$  is the matrix of the observed pretreatment covariates.

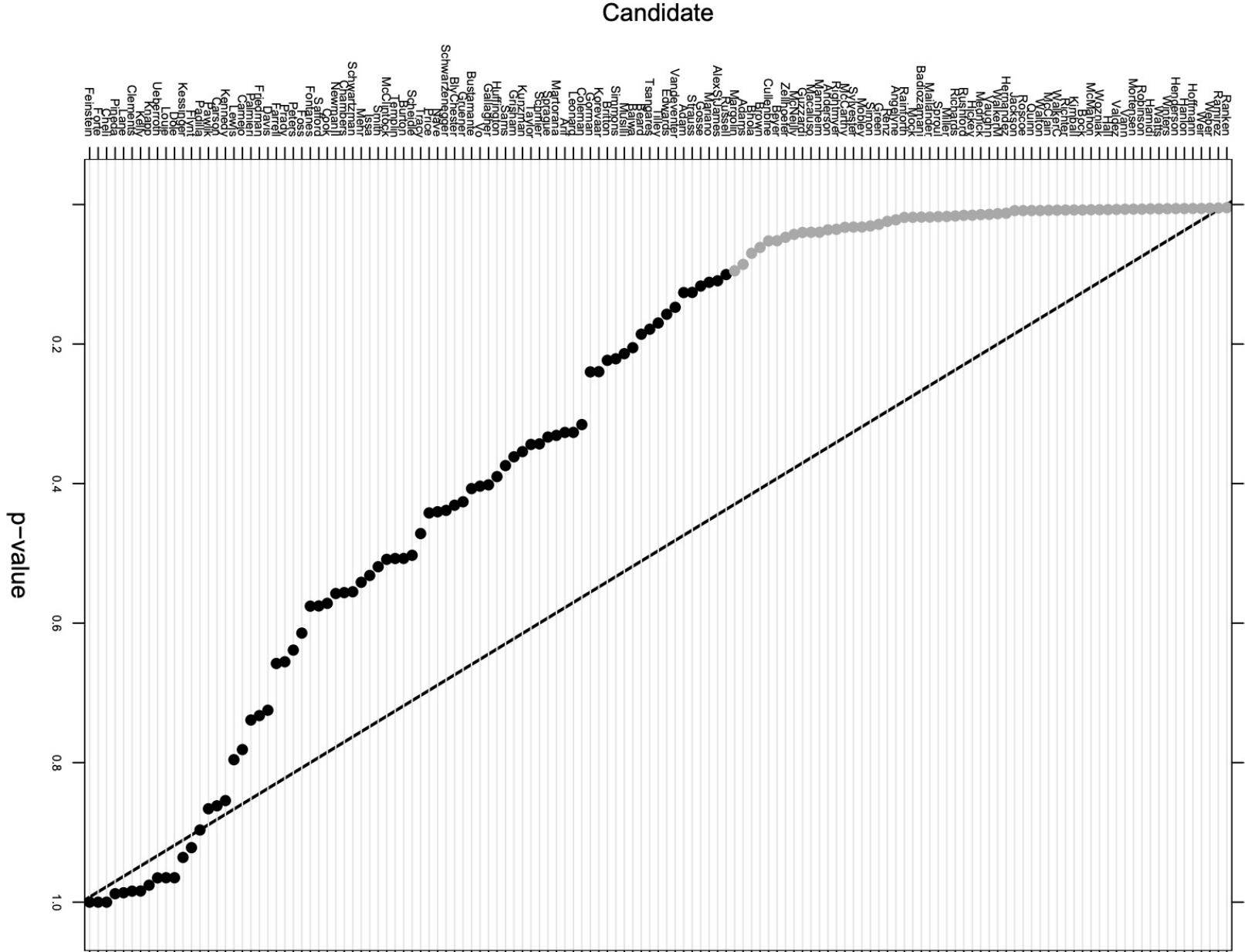
# Set up the analysis framework

## Implicit assumptions

- *Assumption 1* (No interference among units) The potential outcomes of one unit do not depend on the treatment of other units.
  - potential vote shares of a candidate in one district do not depend on the same candidate's ballot placement in another district.
  - Voters usually do not see ballots of other districts and hence are unlikely to be affected by such ballots.
  - focus on the estimation of a separate causal effect for each candidate
- *Assumption 2* (Known random assignment). Treatment is randomly assigned by a known mechanism. Formally,  $p(T_i | Y_i(0), Y_i(1)) = p(T_i)$  is known for each  $i$ .
  - Assumes county page formats are independent of the randomized alphabet
  - Number of possible ballot pages is driven primarily by the type of voting technology, which is exogenous to the randomization

# Distribution of Exact p-values across Candidates

- Authors computed the one-sided p-values
  - Reference distribution obtained via Monte Carlo
  - Candidates ranked based on their p-values
  - If the sharp null is true, these p-values should all be uniformly distributed



# Confidence intervals under the constant additive effect model

- For each candidate, we assume  $Y_i(0) - Y_i(1) \equiv \tau_0$  across all republican / democratic districts
- We construct confidence intervals by inverting the Fisher's randomization tests at a range of  $\tau_0$  values

