Permutation Test

Kosuke Imai

Harvard University

STAT186/GOV2002 CAUSAL INFERENCE

Fall 2018

Randomized Controlled Trials (RTCs)

- Why randomize treatment assignment in experiments?
  1. makes the treatment and control groups "identical"
     - Joint distribution of all observed **X** and unobserved **U** pretreatment confounders is identical:

$$P(\mathbf{X}, \mathbf{U} \mid T = 1) \;=\; P(\mathbf{X}, \mathbf{U} \mid T = 0)$$

     - **U** includes potential outcomes $\{Y(1), Y(0)\}$
     - Treatment assignment is statistically independent of **X** and **U**

$$\{\mathbf{X}, \mathbf{U}\} \perp\!\!\!\perp T \quad \text{and in particular} \quad \{Y(1), Y(0)\} \perp\!\!\!\perp T$$

     - Removes selection problem stochastically $\leftrightsquigarrow$ controlled experiments
  2. enables us to formally quantify the degree of uncertainty

- Potential problems of RCTs: sample selection, placebo effects, noncompliance, missing data, spillover/carryover effects $\rightsquigarrow$ roles of statistical methods

Randomization Inference vs. Model-based Inference

- Randomization as the "reason basis for inference" (Fisher)
- Randomness comes from the physical act of randomization, which then can be used to make statistical inference
- Also called design-based inference
- Advantage: design justifies analysis

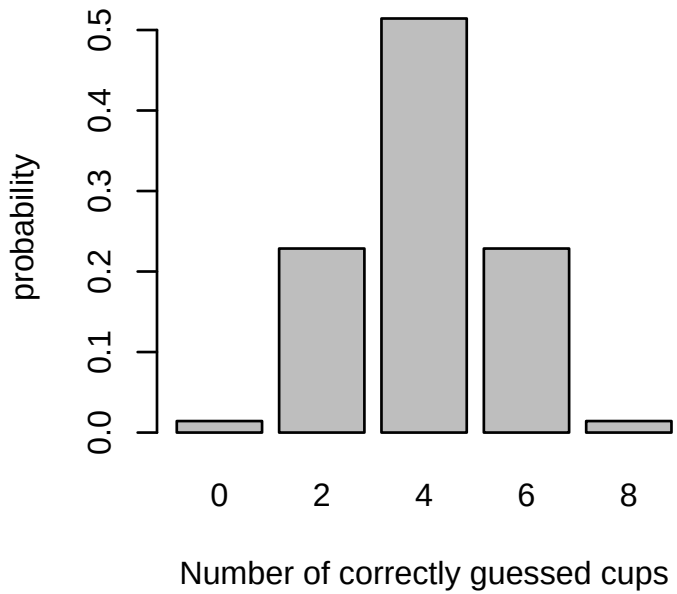
- Contrast this with model-based inference, which assumes a distribution for potential outcomes
- Advantage of model-based inference: flexibility

Lady Tasting Tea (Fisher 1935. *The Design of Experiments*. Oliver and Boyd)

- Does tea taste different depending on whether the tea was poured into the milk or whether the milk was poured into the tea?
- 8 cups; $n = 8$
- Randomly choose 4 cups into which pour the tea first ($T_i = 1$)
- Null hypothesis: the lady cannot tell the difference
- Sharp null – $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, 8$
- Statistic: the number of correctly classified cups
- The lady classified all 8 cups correctly!
- Did this happen by chance?

Permutation Test

cups	guess	actual	scenarios	...
1	M	M	T	T
2	T	T	T	T
3	T	T	T	T
4	M	M	T	M
5	M	M	M	M
6	T	T	M	M
7	T	T	M	T
8	M	M	M	M
correctly guessed		8	4	6



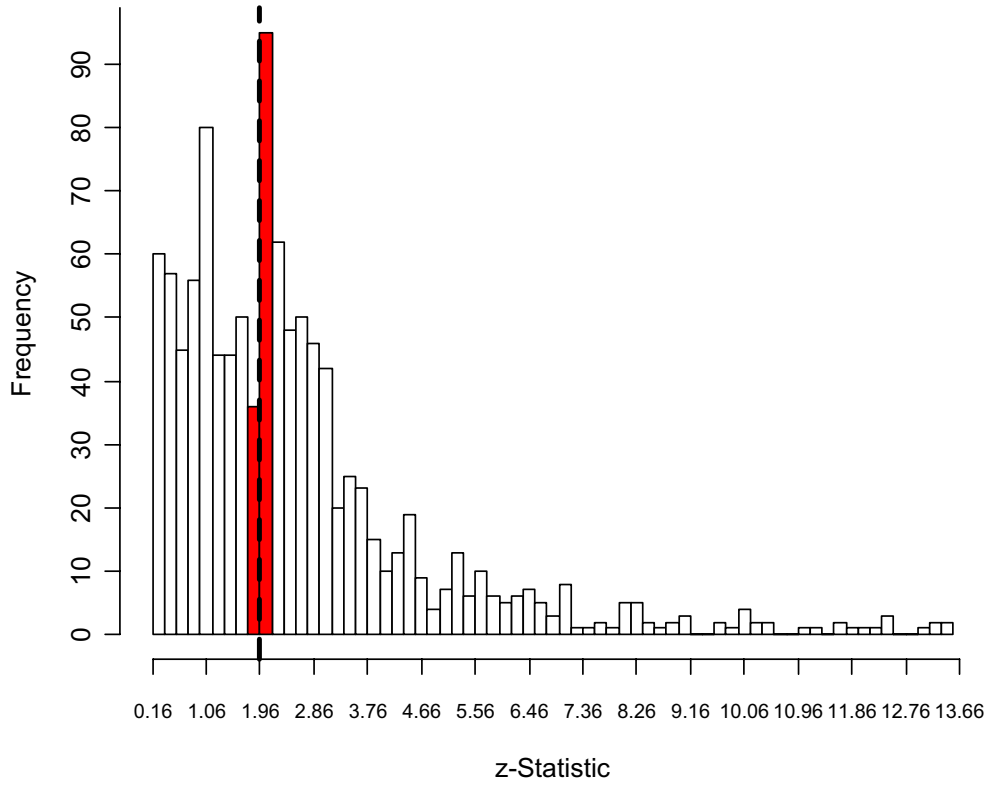
- ${}_8C_4 = 70$ ways to do this and each arrangement is equally likely
- What is the *p*-value?
- No assumption, but the sharp null may be of little interest

5% as the Threshold

It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard.

R. A. Fisher (1935). The Design of Experiments. Oliver & Boyd

- Publication bias:
  - *p*-hacking
  - file drawer bias
- Potential solutions:
  - pre-registration
  - statistical control of multiple testing



(Gerber and Malhotra. 2008. *Q. J. Political Sci.*)

Basic Setup

- Units: $i = 1, \ldots, n$
- Treatment: $T_i \in \{0, 1\}$
- Outcome: $Y_i = Y_i(T_i)$
- Complete randomization of the treatment assignment
  - Exactly n_1 units receive the treatment
  - $n_0 = n - n_1$ units are assigned to the control group
- This differs from the Bernoulli randomization
- We know the distribution of T_i:

$$\Pr(T_i = 1 \mid Y_i(1), Y_i(0)) = \frac{n_1}{n}$$

for all $i = 1, \ldots, n$ and $\sum_{i=1}^{n} T_i = n_1$

Fisher's Exact Test

- 2×2 table:

	Treated ($T = 1$)	Control ($T = 0$)
Success ($Y = 1$)	$\sum_{i=1}^{n} T_i Y_i(1)$	$\sum_{i=1}^{n} (1 - T_i) Y_i(0)$
Failure ($Y = 0$)	$\sum_{i=1}^{n} T_i (1 - Y_i(1))$	$\sum_{i=1}^{n} (1 - T_i)(1 - Y_i(0))$
Total	n_1	n_0

- Test statistic: $S = \sum_{i=1}^{n} T_i Y_i(T_i)$
- Under complete randomization and the sharp null of no effect, the test statistic follows the hyper-geometric distribution:

$$\Pr(S = s) = \frac{\binom{m}{s}\binom{n-m}{n_1-s}}{\binom{n}{n_1}} \quad \text{where } m = \sum_{i=1}^{n} Y_i(T_i)$$

Computation

- Exact computation $\rightsquigarrow$ difficult when n is large
- Monte Carlo approximation:
  1. Fill in missing potential outcomes under the sharp null
  2. Sample T_i according to complete randomization
  3. Compute the test statistic
- Can be made arbitrarily accurate by increasing number of draws
- Analytical approximations:

$$\mathbb{E}(S) = \frac{n_1 m}{n}, \quad \text{and} \quad \mathbb{V}(S) = \frac{m n_1 n_0}{n(n-1)}\left(1 - \frac{m}{n}\right)$$

  1. Normal: $(S - \mathbb{E}(S))/\sqrt{\mathbb{V}(S)} \sim \mathcal{N}(0, 1)$
  2. Binomial$(n_1, m/n)$
- Becomes accurate as n grows

General Procedure for Permutation Tests

1. Specify sharp null hypothesis
   - Typically, $H_0 : \tau_{0i} = Y_i(1) - Y_i(0)$ where we set $\tau_{0i} = 0$ for all i
   - No effect implies no heterogenous effect, no spillover effect, etc.

2. Choose a test statistic $S = f(\{Y_i, T_i, \tau_{0i}\}_{i=1}^N)$
   - Fisher's exact test statistic: $S = \sum_{i=1}^N T_i Y_i(1)$
   - Other commonly used test statistics include rank sum and difference-in-means

3. Compute the reference distribution and p-value based on the randomized distribution of treatment assignment
   - Exact distribution in small samples
   - Large-sample approximation
   - Monte Carlo approximation as a general strategy

Rank-sum Test Statistic

- Fisher's exact test assumes binary outcome
- Rank-sum test is often used for continous outcome
- Wilcoxson's rank-sum statistic:

$$S = \sum_{i=1}^{N} T_i R_i$$

where R_i is the rank statistic

1. symmetric
2. moments (assume no tie):

$$\mathbb{E}(S) = \frac{n_1(n+1)}{2}, \quad \mathbb{V}(S) = \frac{n_0 n_1 (n+1)}{12}$$

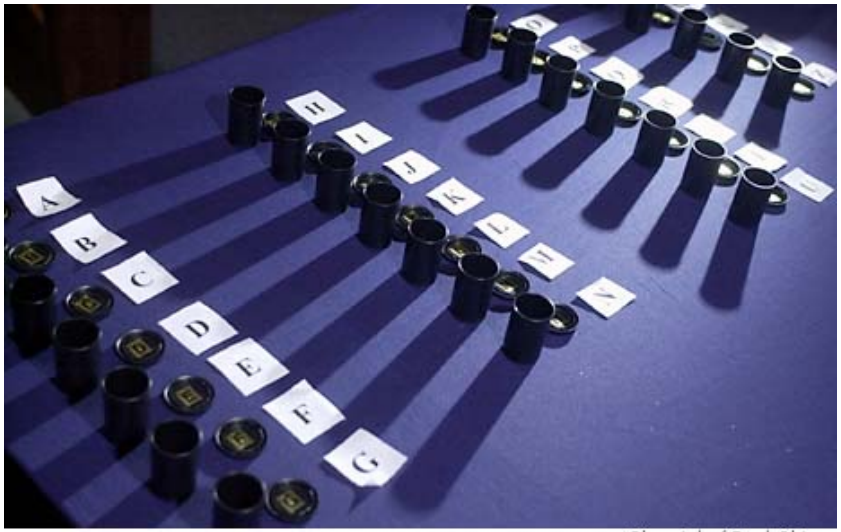
3. reference distribution does not depend on index

- Mann-Whitney U test statistic:

$$U = S - \frac{n_1(n+1)}{2}$$

The California Alphabet Lottery (Ho and Imai, 2006, *J. Am. Stat. Assoc.*)

- Randomization sometimes occurs in the real world
- Started in 1975: "[B]oth the 'incumbent first' and 'alphabetical order' procedures are constitutionally impermissible." Gould v. Grubb, 14 Cal. 3d 661, 676.
- A random alphabet is drawn for every statewide election that applies to all statewide offices



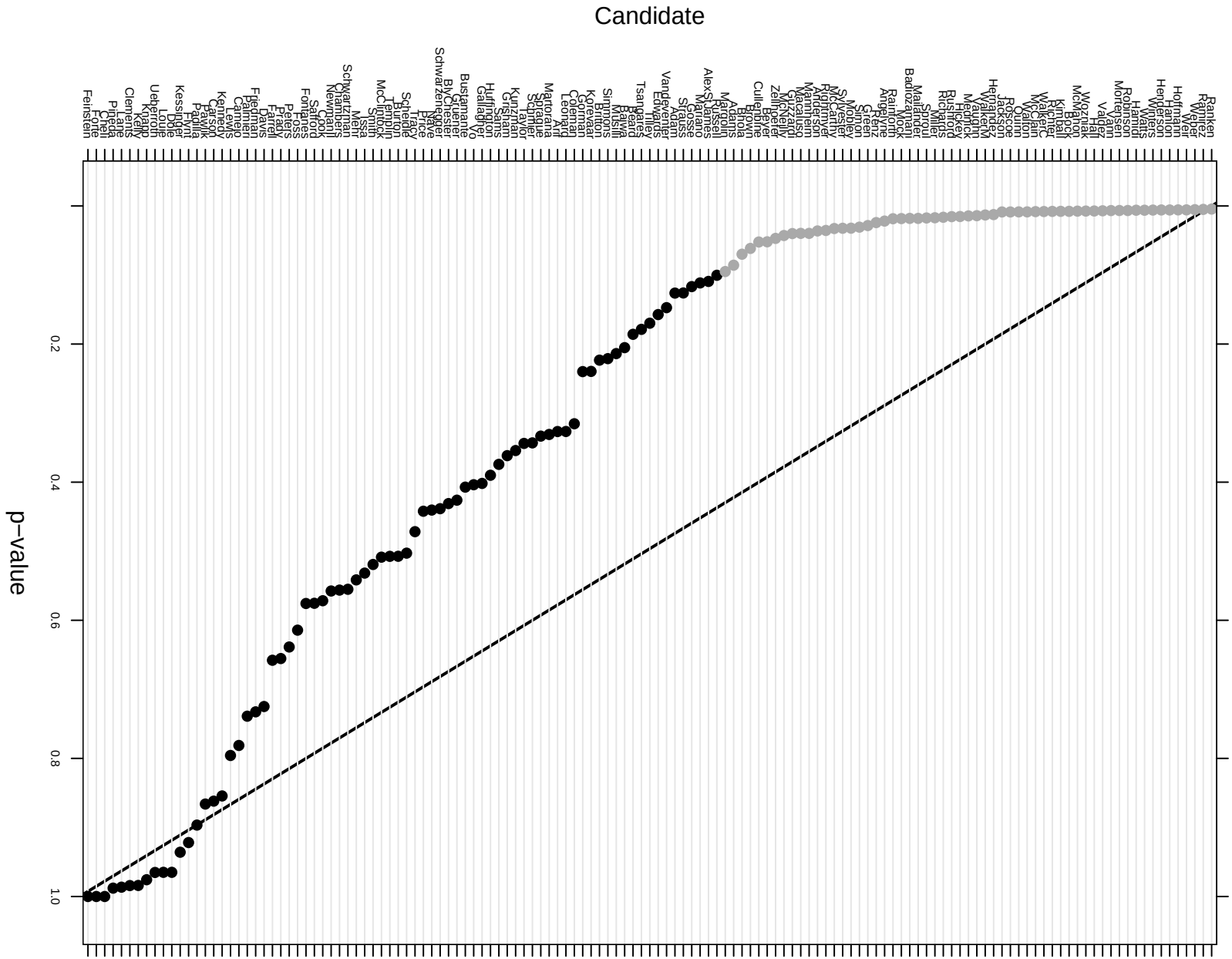
California Elections Code 13112(a)

Each letter of the alphabet shall be written on a separate slip of paper, each of which shall be folded and inserted into a capsule. Each capsule shall be opaque and of uniform weight, color, size, shape, and texture. The capsules shall be placed in a container, which shall be shaken vigorously in order to mix the capsules thoroughly. The container then shall be opened and the capsules removed at random one at a time. As each is removed, it shall be opened and the letter on the slip of paper read aloud and written down. The resulting random order of letters constitutes the randomized alphabet, which is to be used in the same manner as the conventional alphabet in determining the order of all candidates in all elections. For example, if two candidates with the surnames Campbell and Carlson are running for the same office, their order on the ballot will depend on the order in which the letters M and R were drawn in the randomized alphabet drawing.

Permutation Test for the Natural Experiment

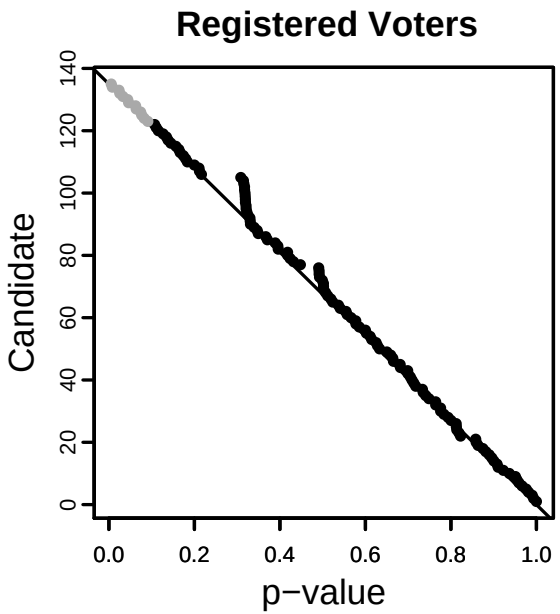
- Take into account the complex lottery procedure
  1. Randomize alphabet
  2. Sort candidates by randomized alphabet
  3. Rotate the candidate order from the first district
- Impossible via model-based inference
- The 2003 CA Gubernatorial Recall Election
- 135 candidates including a porn star
- Ballot order differs across 80 districts
- Null hypothesis: no causal effect of ballot order
- Permutation test for each candidate
- Test statistic: difference-in-means between being on the first ballot page and being on other ballot page
- Reference distribution via Monte Carlo

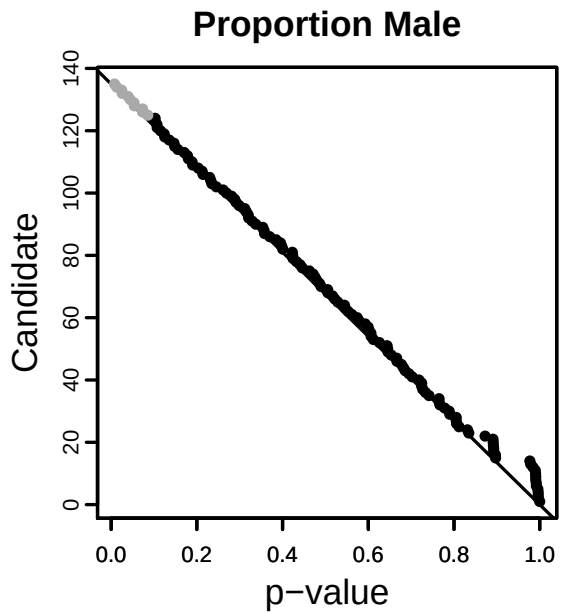
Distribution of Exact *p*-values across Candidates

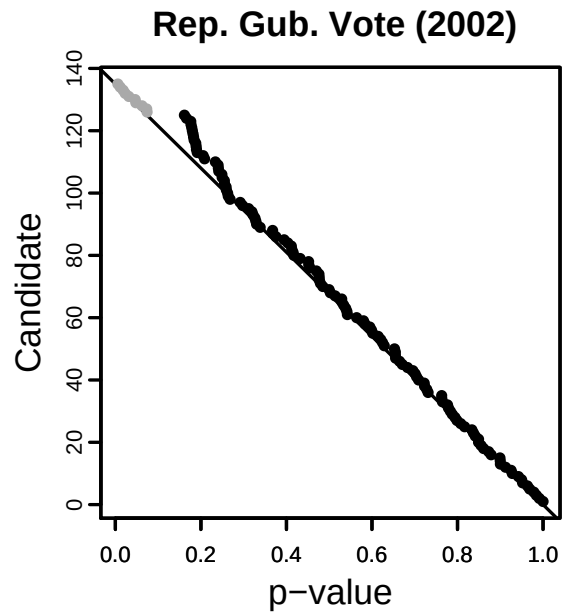


Placebo Tests

- Placebo tests: used when effects are known to be zero
- Null hypothesis is assumed to be true
- Ballot order should not affect pre-election covariates







Summary

- Physical randomization of treatment assignment as a reason basis for inference
- Inference over repeated (hypothetical) randomization

- Advantages:
  1. design-based, assumption-free inference
  2. ability to incorporate complex randomization scheme
- Limitations:
  1. sharp null hypothesis, constant treatment effect
  2. sample inference rather than population inference

- Suggested reading: IMBENS AND RUBIN, CHAPTER 5