Permutation Test

Kosuke Imai

Harvard University

Spring 2021

Agenda

1. Randomized controlled trials
2. Fisher's exact test
3. Rank sum test
4. General permutation tests

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

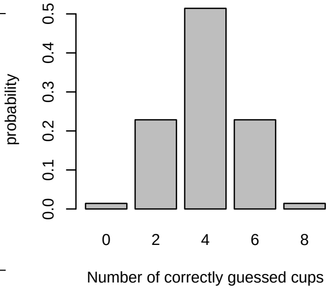
- Does tea taste different depending on whether the tea was poured into the milk or whether the milk was poured into the tea?

- Experiment:
  - Units: 8 identical cups
  - Randomization: Randomly choose 4 cups into which the tea is poured first, and for the other four, the milk was poured first
  - Null hypothesis: the lady cannot tell the difference
  - Statistic: the number of correctly classified cups

- Outcome: The lady classified all 8 cups correctly!
- Did this happen by chance?

Permutation Test

cups	guess	actual	scenarios			
1	M	M	T	T	T	
2	T	T	T	T	M	
3	T	T	T	T	T	
4	M	M	T	M	T	...
5	M	M	M	M	M	
6	T	T	M	M	M	
7	T	T	M	T	M	
8	M	M	M	M	T	
# of correct guesses		8	4	6	2	$\cdots$



- ${}_8C_4 = 70$ ways to do this and each arrangement is equally likely
- Under the null hypothesis, the probability that the lady classifies all cups correctly is $1/70 \approx 0.014$

- The lady may have possessed an ability to tell the difference

Randomized Controlled Trials (RCTs)

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

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

       where **U** includes potential outcomes $\{Y(1), Y(0)\}$
     - Unconfoundedness of treatment assignment:

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

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

- Potential problems of RCTs
  - external validity: sample selection, generalizability
  - human behavior: Hawthorne effect, noncompliance, missing data

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 of potential outcomes
- Advantage of model-based inference: flexibility

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
  - Different from Bernoulli randomization: n_1 and n_0 are not fixed
  - The randomization distribution of T_i:
  $$\Pr(T_i = 1 \mid \mathcal{O}_n) = \frac{n_1}{n} \quad \text{for all } i \text{ and } \sum_{i=1}^{n} T_i = n_1$$
  where $\mathcal{O}_n = \{Y_i(0), Y_i(1)\}_{i=1}^{n}$

- Sharp null hypothesis of no treatment effect:
$$H_0 : Y_i(1) = Y_i(0) \quad \text{for all } i.$$

Fisher's Exact Test

- 2×2 table:

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

- Test statistic (# of successes in treatment group):

$$S = \sum_{i=1}^n T_i Y_i(1)$$

- Under the sharp null, we have $Y_i(1) = Y_i(0) = Y_i$
- Reference distribution (hyper-geometric distribution):

$$\Pr(S = s \mid \mathcal{O}_n) = \frac{\begin{array}{c}\text{assign } s \text{ successes}\\ \text{to treatment group}\end{array} \times \begin{array}{c}\text{assign } n_1 - s \text{ failures}\\ \text{to treatment group}\end{array}}{\begin{array}{c}\text{\# of ways to assign } n_1\\ \text{units to treatment group}\end{array}} = \frac{\binom{m}{s}\binom{n-m}{n_1-s}}{\binom{n}{n_1}}$$

Computation

- Exact computation $\rightsquigarrow$ difficult when n is large
- Analytical approximations:

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

  1. Normal: $\{S - \mathbb{E}(S \mid \mathcal{O}_n)\}/\sqrt{\mathbb{V}(S \mid \mathcal{O}_n)} \sim \mathcal{N}(0,1)$
  2. Binomial$(n_1, m/n)$
- Becomes accurate as n grows
- 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
- Widely applicable so long as the treatment assignment mechanism is known

The Project STAR (Mosteller. 1997. *Bull. Am. Acad. Arts Sci.*)

- The Student-Teacher Achievement Ratio Project (1985–1989)
  - More than 10,000 students involved with the cost of $12 million
  - Effects of class size in early grade levels
  - 3 arms: Small class, Regular-sized class, Regular class with aid

- Long-term impact of class size:

	Small class	Regular-sized class
Graduate	754	892
Not graduate	148	189
Total	902	1081

  - Exact *p*-value: 0.28 (one-sided), 0.55 (two-sided)
  - Asymptotic *p*-value: 0.26 (one-sided), 0.53 (two-sided)

Rank-sum Tests

- Fisher's exact test assumes binary outcome
- Rank-sum tests are often used for continuous outcome
- Rank of the outcome for unit i: $R_i = R_i(Y_1(T_1), \ldots, Y_n(T_n))$
- Wilcoxon's rank-sum statistic:

$$S = \sum_{i=1}^{n} T_i R_i(Y_1(T_1), \ldots, Y_n(T_n))$$

  1. symmetric around the mean $\rightsquigarrow$ good for normal approximation
  2. moments (under the assumption of no tie):

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

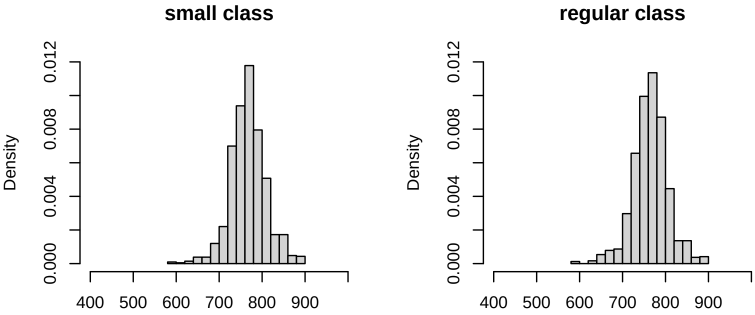
  3. reference distribution does not depend on scale and is not sensitive to outlier

- Mann-Whitney U test statistic (mean zero):

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

The Project STAR Revisited

- Effect of kindergraden class size on 8th grade reading score:



- Wilcoxon's rank-sum test (there are some ties):
  p-value ≈ 0.14

General Procedure for Permutation Tests

1. Specify a 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, rank sum test statistic, etc.
   - Any statistic gives a valid and exact p-value but power may differ
   - Could use regression models or machine learning algorithms

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

Summary

- Randomization of treatment assignment as a “reason basis for inference” $\rightsquigarrow$ design-based, assumption-free inference
- Inference over repeated (hypothetical) randomization $\rightsquigarrow$ sample inference rather than population inference

- Sharp null hypothesis:
  - implies no effect for every unit
  - may not be of interest but serves as a starting point of analysis

- Permutation test as a general testing procedure for RCTs
  - flexibility: any test statistic can be used with Monte Carlo simulation
  - assumption-free