# Lecture 3
# randomized experiments and Fisher’s exact p-value

# Outline

- Five examples of randomized experiment mechanisms
- Fisher's exact p-value
  - Fisher's original experiment
  - Hypothesis testing
  - Construct confidence intervals
  - Choice of the test statistics
- Suggested reading: Imbens and Rubin Chapter Chapter 4-5, Peng's book Chapter Chapter 3

# Treatment assignment mechanism

- Assignment vector for binary treatment with N units: $\boldsymbol{W} = (W_1, \cdots, W_N) \in \{0,1\}^N$

- Unconfoundedness property: $P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = P(\boldsymbol{W}|\boldsymbol{X})$
  - Assignment mechanism does not depend unobserved $\boldsymbol{U}$ pretreatment confounders
  - $\boldsymbol{U}$ includes potential outcomes $\left(Y_i(0), Y_i(1)\right)$
  - We can alternatively understand it as

$$W_i \perp \left(Y_i(0), Y_i(1)\right) \mid \boldsymbol{X}_i$$

  - Make the treatment and control groups "identical"

$$P(Y_i(0), Y_i(1)| \boldsymbol{X}_i, W_i = 0) = P(Y_i(0), Y_i(1)| \boldsymbol{X}_i, W_i = 1)$$

- Identify conditional population average treatment effect under unconfoundedness

Causal estimand that involve the unobserved potential outcomes

$$\begin{aligned}\tau(\boldsymbol{x}) &= \mathbb{E}(Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}) \\ &= \mathbb{E}(Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1) - \mathbb{E}(Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = 0) \\ &= \mathbb{E}(Y_i(W_i)| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1) - \mathbb{E}(Y_i(W_i) \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = 0) \\ &= \mathbb{E}(Y_i| \boldsymbol{X}_i = \boldsymbol{x}, W_i = 1) - \mathbb{E}(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}, W_i = 0)\end{aligned}$$

Conditional expectations that we can evaluate based on observed data

# Common designs of randomized experiments

- Five examples of randomized experiment mechanisms
  - Bernoulli trial
  - Completely randomized experiment
  - Stratified randomized experiment
  - Paired randomized experiment
  - Rerandomization

- The purpose of restricting the assignment mechanism is to eliminate assignment vectors that are less desirable for estimating causal effects
  - Examples: all males get treatment; all females get control

# Bernoulli trial

- Simplest Bernoulli experiment tosses a (fair) coin for each unit
  - If the coin is heads, then unit receive treatment
  - Otherwise, the unit receive control

- For each $\boldsymbol{w} \in \{0,1\}^N$, $P(\boldsymbol{W} = \boldsymbol{w} | \boldsymbol{X}) = P(\boldsymbol{W} = \boldsymbol{w}) = 0.5^N$
- $W_1, \cdots, W_N \sim \text{Bernoulli}(0.5)$ and are independent

- More generally, we can toss a specialized coin for each unit depending on its covariates
  - Define propensity score $e(\boldsymbol{X}_i) = P(W_i = 1 | \boldsymbol{X}_i)$
  - Assignment property: $P(\boldsymbol{W} = \boldsymbol{w} | \boldsymbol{X}) = \prod_{i=1}^{N} [e(\boldsymbol{X}_i)^{W_i} (1 - e(\boldsymbol{X}_i))^{1-W_i}]$
  - $W_1, \cdots, W_N$ are still independent and each $W_i \sim \text{Bernoulli}(e(\boldsymbol{X}_i))$
  - Example: when trying to induce people with serious disease to enroll for the trial of a promising drug, we give them a higher probability to receive the treatment

- Drawback of the design: always a positive probability that all units receive the same treatment

# Completely randomized experiment

- A fixed number of subjects $N_t$ is assigned to receive the active treatment

- Assignment probability

$$P(\boldsymbol{W} = \boldsymbol{w} | \boldsymbol{X}) = \begin{cases} \binom{N}{N_t}^{-1} & \text{if } \sum_{i=1}^{N} w_i = N_t \\ 0 & \text{otherwise} \end{cases}$$

- Completely randomized experiment guarantees that there are exactly $N_t$ individuals receiving the treatment and $N - N_t$ individuals receiving the control
- $W_1, \cdots, W_N$ are slightly negatively associated

- There is still positive probability that all females receive the control and all females receive treatment → extreme covariate imbalance after randomization
- In that case, average differences between groups could be due to sex differences
- For this single experiment, we can get a terrible estimate and wrong judgement

# Stratified randomized experiment

- Basic procedure:
  1. Blocking (Stratification): create groups of similar units based on pre-treatment covariates, let $B_i \in \{1, \cdots, J\}$ be the block indicator
  2. Block (Stratified) randomization: completely randomize treatment assignment within each group
- Blocking can improve the efficiency by minimizing the variance of the potential outcomes within each strata

*"Block what you can and randomize what you cannot"*

Box, et al. (2005). Statistics for Experimenters. 2nd eds. Wiley

- Assignment probability

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

# Examples

- Randomized trial for the Moderna vaccine
  [Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England journal of medicine*, 2020.]
  - Participants were randomly assigned in a 1:1 ratio, through the use of a centralized interactive response technology system, to receive vaccine or placebo.
  - Assignment was stratified into the following risk groups: persons 65 years of age or older, persons younger than 65 years of age who were at heightened risk (at risk) for severe Covid-19, and persons younger than 65 years of age without heightened risk (not at risk).

- Experiment of women policy makers in India
  [Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica,*2004]
  - Each Gram Panchayat (GP) encompasses 10,000 people in several villages (between 5 and 15)
  - Starting 1993, in a third of the villages in each GP, only women could be candidates for the position of councilor for the area.
  - Random selection: villages are ranked in consecutive order according to an administrative number, every third village is reserved for a woman
  - How is the experiment stratified?

# Paired randomized experiment

- Can we keep blocking until we cannot block any further?
- Procedure:
  1. Create $J = N/2$ pairs of similar units
  2. Randomize treatment assignment within each pair

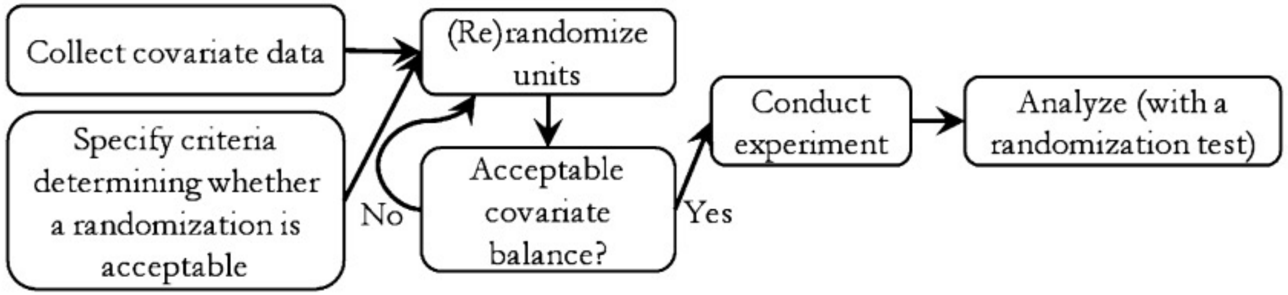
- Example: evaluation of health insurance policy
  [Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. *The lancet, 2009*.]
  - Seguro Popular, a programme aimed to deliver health insurance, regular and preventive medical care, medicines, and health facilities to 50 million uninsured Mexicans
  - Units: health clusters = predefined health facility catchment areas
  - Randomization within 74 matched pairs of “similar” health clusters
  - Outcome: proportion of households within each health cluster who experienced catastrophic medical expenditure

# Rerandomization

[Morgan and Rubin. 2012. Ann. Stat., Li et al. 2018. PNAS]

- The more covariates, the more likely at least one covariate will be imbalanced across treatment groups
- Randomization only eliminate confounding factors and yield unbiased on average (over repeated run of experiments)
- For any particular experiment, covariate imbalance is possible

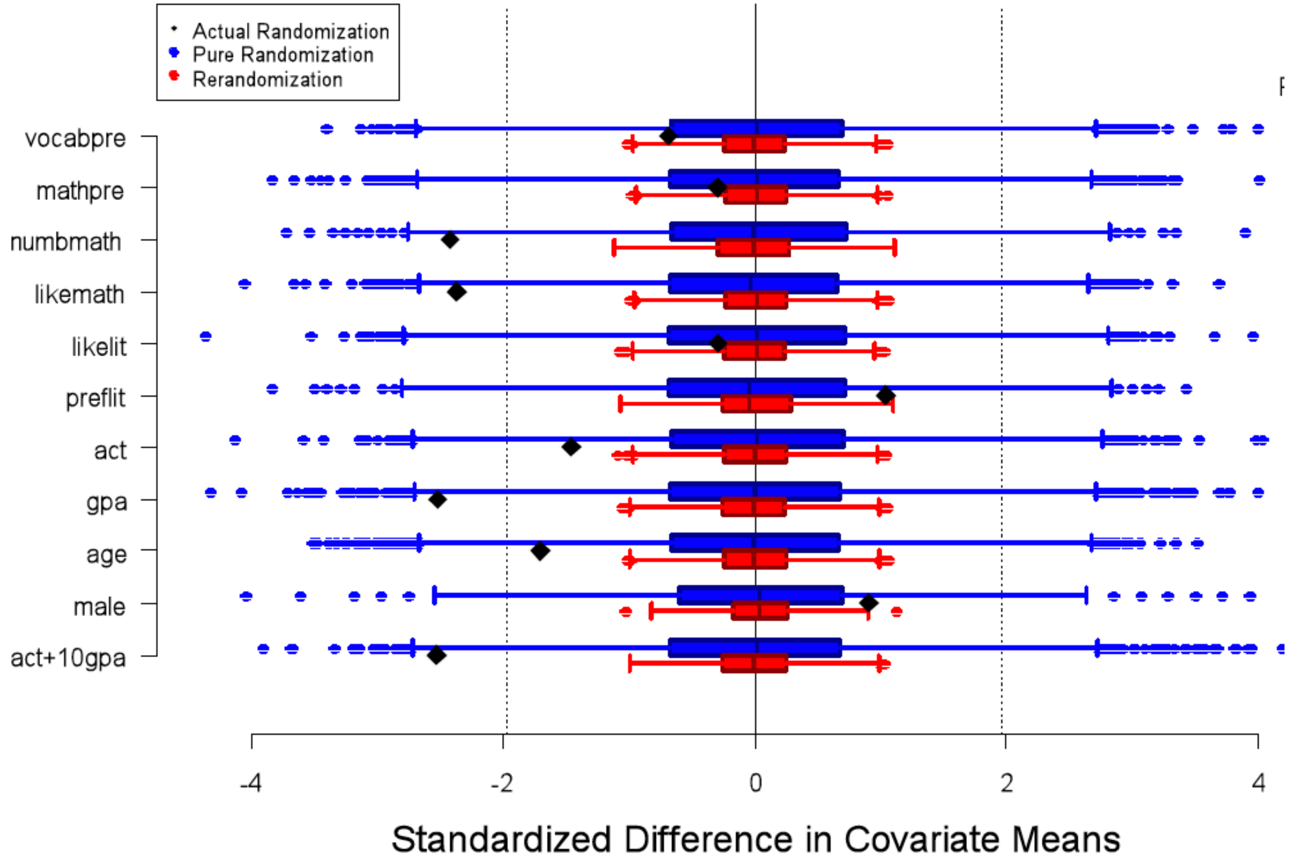
- Procedure:
  1. Specify the acceptable level of covariate balance
  2. Randomize the treatment and check covariate balance
  3. Repeat until the covariate balance criterion is met



# Rerandomization: an example

[Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, 2015.]

- The study aim to examine whether observational studies can be analyzed to yield valid estimates of causal effects.
- Undergraduate psychology students at a particular college were randomized to be in one of two arms: a randomized experiment ($n_r$ = 235) or an observational study ($n_o$ = 210).
- In the randomized experiment were randomized to take either a vocabulary or mathematics course



# 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

- model-based inference: assume a distribution for potential outcomes (at least the i.i.d. assumptions)
- Advantage of model-based inference: flexibility

- Two types of classical randomization inference
  - Fisher’s exact p-values
  - Neyman’s repeated sampling approach

# Fisher's original experiment: Lady tasting tea

[Fisher, 1935]

- The lady in question (Muriel Bristol) claimed to be able to tell whether the tea or the milk was added first to a cup.
- Fisher proposed to give her eight cups, four of each variety, in random order.

- **Null hypothesis:** the lady cannot tell the difference
- How to define the causal effect?
  - whether the tea or the milk was added first has any effect on the lady's guess result
- What is a unit?
- What is the treatment assignment?
  - $W_i = 1$ if tea is added first
- What is $Y_i(0)$ and $Y_i(1)$?
  - The lady's potential guess results

- **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, 8$

# Fisher's original experiment: Lady tasting tea

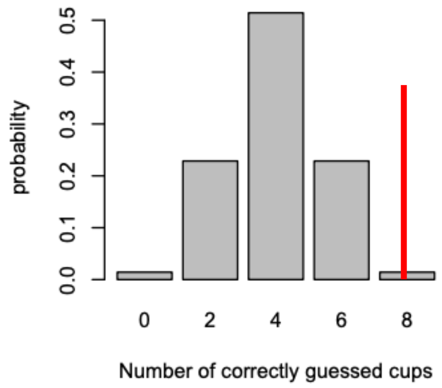
[Fisher, 1935]

- **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, 8$
- **Test statistics:** the number of correctly classified cups
- The lady classified all 8 cups correctly! Did this happen by chance?
- Goal: calculate the distribution of the test statistics under the null

- Completely randomized experiment: $\binom{8}{4} = 70$ possible scenarios with equal probability
- Under the sharp null, the lady will always have the same guesses under all scenarios

| | $Y_i(0) = Y_i(1) = Y_i$ | $W_i$ | 70 possible assignments | | |
|---|---|---|---|---|---|
| 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 | |

Null distribution of the test statistics




P-value: 1/70

# Cough frequency example with 6 units

**Table 5.3.** ***Cough Frequency for the First Six Units from the Honey Study***

| Unit | Potential Outcomes | | | | |
|---|---|---|---|---|---|
| | Cough Frequency (cfa) | | Observed Variables | | |
| | $Y_i(0)$ | $Y_i(1)$ | $W_i$ | $X_i$ (cfp) | $Y_i^{obs}$ (cfa) |
| 1 | ? | 3 | 1 | 4 | 3 |
| 2 | ? | 5 | 1 | 6 | 5 |
| 3 | ? | 0 | 1 | 4 | 0 |
| 4 | 4 | ? | 0 | 4 | 4 |
| 5 | 0 | ? | 0 | 1 | 0 |
| 6 | 1 | ? | 0 | 5 | 1 |

Imputation under the sharp null

| | Cough Frequency (cfa) | | | |
|---|---|---|---|---|
| | $Y_i(0)$ | $Y_i(1)$ | $Y_i^{obs}$ | $\text{rank}(Y_i^{obs})$ |
| 1 | (3) | 3 | 3 | 4 |
| 2 | (5) | 5 | 5 | 6 |
| 3 | (0) | 0 | 0 | 1.5 |
| 4 | 4 | (4) | 4 | 5 |
| 5 | 0 | (0) | 0 | 1.5 |
| 6 | 1 | (1) | 1 | 3 |

- **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, 6$
  absolutely no causal effect of the treatment
- Test statistics: $|\bar{Y}_t^{obs} - \bar{Y}_c^{obs}|$ or $|\overline{\text{rank}}_t(Y_i^{obs}) - \overline{\text{rank}}_c(Y_i^{obs})|$

# Cough frequency example with 6 units

- If following completely randomized experiment: $\binom{6}{3} = 20$ assignments with equal probability

**Table 5.5.** ***Randomization Distribution for Two Statistics for the Honey Data from Table 5.3***

| | | | | | | Statistic: Absolute Value of Difference in Average | |
|---|---|---|---|---|---|---|---|
| $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ | $W_6$ | Levels ($Y_i$) | Ranks ($R_i$) |
| 0 | 0 | 0 | 1 | 1 | 1 | −1.00 | −0.67 |
| 0 | 0 | 1 | 0 | 1 | 1 | −3.67 | −3.00 |
| 0 | 0 | 1 | 1 | 0 | 1 | −1.00 | −0.67 |
| 0 | 0 | 1 | 1 | 1 | 0 | −1.67 | −1.67 |
| 0 | 1 | 0 | 0 | 1 | 1 | −0.33 | 0.00 |
| 0 | 1 | 0 | 1 | 0 | 1 | 2.33 | 2.33 |
| 0 | 1 | 0 | 1 | 1 | 0 | 1.67 | 1.33 |
| 0 | 1 | 1 | 0 | 0 | 1 | −0.33 | 0.00 |
| 0 | 1 | 1 | 0 | 1 | 0 | −1.00 | −1.00 |
| 0 | 1 | 1 | 1 | 0 | 0 | 1.67 | 1.33 |
| 1 | 0 | 0 | 0 | 1 | 1 | −1.67 | −1.33 |
| 1 | 0 | 0 | 1 | 0 | 1 | 1.00 | 1.00 |
| 1 | 0 | 0 | 1 | 1 | 0 | 0.33 | 0.00 |
| 1 | 0 | 1 | 0 | 0 | 1 | −1.67 | −1.33 |
| 1 | 0 | 1 | 0 | 1 | 0 | −2.33 | −2.33 |
| 1 | 0 | 1 | 1 | 0 | 0 | 0.33 | 0.00 |
| 1 | 1 | 0 | 0 | 0 | 1 | 1.67 | 1.67 |
| 1 | 1 | 0 | 0 | 1 | 0 | 1.00 | 0.67 |
| 1 | 1 | 0 | 1 | 0 | 0 | 3.67 | 3.00 |
| 1 | 1 | 1 | 0 | 0 | 0 | **1.00** | **0.67** |

- P-value based on test statistics $\left|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right|$: $\frac{16}{20} = 0.8$
- P-value based on test statistics $|\overline{\text{rank}}(Y_t^{\text{obs}}) - \overline{\text{rank}}(Y_c^{\text{obs}})|$: $\frac{16}{20} = 0.8$

- The most extreme p-value we can get: 2/20 = 0.1
- $N = 6$ is too small to obtain statistically significant rejections

# Illustration of Fisher's randomization test

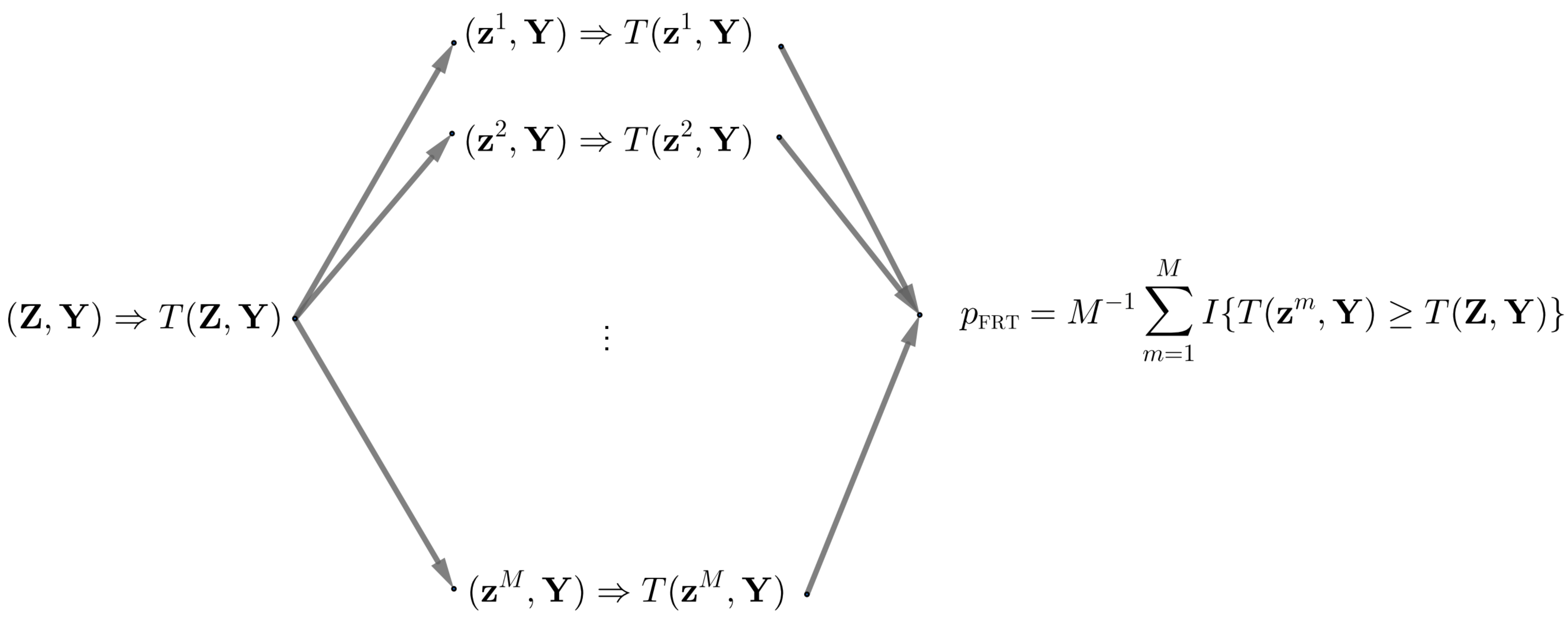



Figure 3.1 of Peng's book

# Fisher's exact p-value

- Features
  - Justified by randomization alone: No assumptions about models or asymptotic normality
  - The sharp null may be of little interest
  - P-value is exact for small $N$
  - Same idea as a permutation test

- Computation of p-value
  - Exact computation is difficult when $N$ is large
  - Monte Carlo approximation
    1. Fill in missing potential outcomes under the sharp null
    2. Sample $W_i$ according to complete randomization
    3. Compute the test statistic to form a reference distribution
  - Approximation can be arbitrarily accurate by increasing number of draws

- Fisher's exact p-value can be calculated for any randomization mechanism
- Analytical approximations when $N$ is large (omitted)

# Cough frequency example with $N = 72$

P-value computation with Monte Carlo approximation

| Number of Simulations | P-Value | $\widehat{(\text{s. e.})}$ |
|---|---|---|
| 100 | 0.010 | (0.010) |
| 1,000 | 0.044 | (0.006) |
| 10,000 | 0.044 | (0.002) |
| 100,000 | 0.042 | (0.001) |
| 1,000,000 | 0.043 | (0.000) |

*Note*: Statistic is absolute value of difference in average ranks of treated and control cough frequencies. P-value is proportion of draws at least as large as observed statistic.

# Fisher's exact p-value and CI

- Choice of the null hypothesis
  - Sharp null of no treatment effect: $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, N$
  - Fisher's approach cannot accommodate a null hypothesis of **zero average effect :** we can not impute the unmeasured potential outcomes

  - Allow more general null hypothesis $H_0: Y_i(0) = Y_i(1) + C_i$ for all $i = 1, \cdots, N$ with pre-defined $(C_1, \cdots C_N)$

- Invert Fisher's exact p-values for confidence intervals of $\tau_0$
  - Assume the constant additive effect model $Y_i(0) - Y_i(1) \equiv \tau_0$
  - We can still impute the missing potential outcomes under the above null with a pre-specified $\tau_0$
  - Collect all null values $\tau_0$ that cannot be rejected by $\alpha$-level Fisher's exact test
  - Idea: if we cannot reject a null hypothesis with a particular effect size, then the confidence interval should include it

# Cough frequency example revisited

- We want to test for the generalized sharp null $H_0 : Y_i(1) - Y_i(0) \equiv 0.5$
  - We need to impute the missing values differently under the new $H_0$
  - The test statistics is different
    - Based on the mean difference $|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} - 0.5|$
    - Based on the rank: we define the rank of each unit based on $\text{rank}(Y_i(0))$[or equivalently $\text{rank}(Y_i(1))$], instead of $\text{rank}(Y_i^{\text{obs}})$

| Unit | Potential Outcomes | | Actual Treatment | Observed Outcome |
|---|---|---|---|---|
| | $Y_i(0)$ | $Y_i(1)$ | | |
| 1 | (2.5) | 3.0 | 1 | 3.0 |
| 2 | (4.5) | 5.0 | 1 | 5.0 |
| 3 | (−0.5) | 0.0 | 1 | 0.0 |
| 4 | 4.0 | (4.5) | 0 | 4.0 |
| 5 | 0.0 | (0.5) | 0 | 0.0 |
| 6 | 1.0 | (1.5) | 0 | 1.0 |

# Cough frequency example with $N = 72$

| Hypothesized Treatment Effect | P-Value (level) | P-Value (rank) |
|---|---|---|
| −3.00 | 0.000 | 0.000 |
| −2.75 | 0.000 | 0.000 |
| −2.50 | 0.000 | 0.000 |
| −2.25 | 0.000 | 0.000 |
| −2.00 | 0.001 | 0.000 |
| −1.75 | 0.006 | 0.078 |
| −1.50 | 0.037 | 0.078 |
| −1.44 | 0.050 | 0.078 |
| −1.25 | 0.146 | 0.078 |
| −1.00 | 0.459 | 0.628 |
| −0.75 | 0.897 | 0.428 |
| −0.50 | 0.604 | 0.428 |
| −0.25 | 0.237 | 0.429 |
| 0.00 | 0.067 | 0.043 |
| 0.06 | 0.050 | 0.043 |
| 0.25 | 0.014 | 0.001 |
| 0.50 | 0.003 | 0.000 |
| 0.75 | 0.000 | 0.001 |
| 1.00 | 0.000 | 0.000 |

- 95% CI based on statistics $\left|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} - \tau_0\right|$:
  - $[-1.44, 0.06]$
- 95% CI based on statistics $\left|\overline{\text{rank}}_t\big(Y_i(0)\big) - \overline{\text{rank}}_c\big(Y_i(0)\big)\right|$:
  - $[-2, 0]$

# Choice of test statistics

- Fisher's exact p-values are valid for any test statistics
  - Choice of test statistic determines "power" to detect a particular alternative hypotheses
  - Choose a test statistics that is sensitive to expected departures from the null hypothesis
    - One principle: test statistics is "centered at 0" under the null

- Examples for test statistics for the sharp null $H_0: Y_i(0) \equiv Y_i(1)$
  - Sample mean difference: $|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}|$
  - Sample rank mean difference: $\left|\overline{\text{rank}}_t\left(Y_i^{\text{obs}}\right) - \overline{\text{rank}}_c\left(Y_i^{\text{obs}}\right)\right|$
    [check Imbens and Rubin book page 57 for a formal definition of a normalized rank with ties]

  - Quantile difference (more robust to outliers)
    - Difference in medians: $|\text{med}_t\left(Y_i^{\text{obs}}\right) - \text{med}_c\left(Y_i^{\text{obs}}\right)|$

  - Fisher's exact test statistics for binary outcome: $S = \sum_{i=1}^{n} W_i Y_i(1) = \sum_{i=1}^{n} W_i Y_i$
  - Covariate-adjusted statistics

# Fisher's exact test for binary outcome

| | Treated ($W_i = 1$) | Control ($W_i = 0$) | Total |
|---|---|---|---|
| $Y_i = 1$ | $\sum_{i=1}^{n} W_i Y_i(1)$ | $\sum_{i=1}^{n} (1 - W_i) Y_i(0)$ | $m$ |
| $Y_i = 0$ | $\sum_{i=1}^{n} W_i (1 - Y_i(1))$ | $\sum_{i=1}^{n} (1 - W_i)(1 - Y_i(0))$ | $N - m$ |
| Total | $N_1$ | $N_0$ | $N$ |

- In the tea tasting example, the lady knows that there are 4 cups for each variety, so $m$ is also fixed
- Test statistics: $S = \sum_{i=1}^{n} W_i Y_i(1) = \sum_{i=1}^{n} W_i Y_i$
- Then under complete randomization and the sharp null, $S$ follows a hyper-geometric distribution

$$P(S = s) = \frac{\binom{m}{s}\binom{N - m}{N_1 - s}}{\binom{N}{N_1}}$$

- Under the sharp null $H_0$: $Y_i(0) \equiv Y_i(1)$, $m$ is always naturally fixed as $Y_i \equiv Y_i(0) \equiv Y_i(1)$are always fixed

# Lady tasting tea revisited: R code

```r
# Data setup: a 2x2 contingency table where rows represent guessed (milk first or tea first),
# and columns represent actual preparation (milk first or tea first).
data <- matrix(c(4, 0, 0, 4), nrow = 2,
               dimnames = list(Guessed = c("Milk First", "Tea First"),
                               Actual = c("Milk First", "Tea First")))
```

```
             Actual
Guessed       Milk First Tea First
  Milk First          4         0
  Tea First           0         4
```

```r
# Perform Fisher's Exact Test
fisher_result <- fisher.test(data, alternative = "greater")
```

```
> print(fisher_result)

        Fisher's Exact Test for Count Data

data:  data
p-value = 0.01429


> 1/70
[1] 0.01428571
```

# The project STAR example

(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 |

- Check by yourself with R
- p-values: 0.28 (one-sided), 0.55 (two-sided)