From identification to estimation

• So far, we have focused on identification of causal effects
• We discussed the identifiability conditions
  • Exchangeability, positivity, consistency
  • That is, we have simply assumed that the probabilities in question are sufficiently accurately estimated
  • The analysis is based on an infinite study population which allowed us to ignore random variability across finite samples from the study population
• This is seldom the case in practice
• We now turn to causal effect estimation from finite samples
• We will also address practical approaches to enforcing the identifiability conditions
Randomized experiments

- The gold standard for causal inference
- **Suppose we want to determine whether the color (blue, black) of the ink used by a student to answer an exam has a causal effect on grade (pass versus fail)**
- Can you think of a suitable experiment?
  - You stand by the door
  - As a student is about to enter the room, you flip a fair coin
    - If the outcome is a head, you give the student a black pen
    - If the outcome is a tail, you give the student a blue pen
  - After the instructor grades the exams, you compare the proportion of students receiving the passing grade among those receiving black pens with that of those receiving blue pens
  - If the two proportions are same, then the color of the pen has no causal effect on grade

Estimating causal effects

- To estimate causal effects, we want treated and untreated groups to be similar with respect covariates
- **Why? To ensure exchangeability**
- We do so by creating covariate balance across treatment groups
- Easiest way to accomplish this: randomized experiments
Randomized experiment

• The assignment mechanism is random, known, and controlled by the researcher

• Because the treatments are randomly assigned, the treatment groups should all look similar regarding covariates (observed and unobserved)

• In such “classical” randomized experiments, the assignment mechanism is unconfounded by design

• If possible, randomized experiments should be double-blind: neither the participants or the researchers involved should know which treatment the patients are actually getting
Analysis of RCT under the exchangeability assumption

- **Assignment to Blue and Black groups is randomized**
- The proportion of “Pass”, i.e., outcome 1, among the Black group is expected to be identical to those in the Blue group had it been the case that the Blue group were treated (received Black pens) instead of the Black group
- **The treated and untreated groups are exchangeable**

<table>
<thead>
<tr>
<th>Person</th>
<th>W</th>
<th>Y_{A=1}</th>
<th>Y_{A=0}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>

Assignment vector

- **Exchangeability** means that the expected outcome will be same in the two groups if both received treatment or both did not receive treatment
- **In other words, the treated and untreated groups are similar with respect to the covariates**

\[
\text{Causal effect of treatment} = \Pr[Y_{A=1} = 1] - \Pr[Y_{A=0} = 1] = \Pr(Y = 1|A = 1) - \Pr(Y = 1|A = 0) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}
\]
Randomized control trial

- Randomization is expected to ensure exchangeability
- Hence, in the case of RCT, causation is association!
- Counterfactual outcome $Y^a$ is statistically independent of the observed treatment $A$, i.e., $Y^a \perp \! \! \! \! \! \! \! \! \! \! \! \perp A$
- Independence between counterfactual outcome and the observed treatment does not imply independence between the observed outcome and observed treatment
- When the treatment has a causal effect on the outcome, $Y$ and $A$ are associated and hence not independent!
- Experiments where exchangeability holds are called marginally randomized experiments

Randomized Experiments

- Four types of classical randomized experiments:
  - Bernoulli randomized experiment
  - Completely randomized experiment
  - Stratified randomized experiment
  - Paired randomized experiment
- Increasingly restrictive treatment assignments
Bernoulli

- In a Bernoulli experiment, the treatment for each unit is determined by a coin flip
- Treatment assignments for units are independent
- Usually, $e(x) = \frac{1}{2}$
  - $e(x) = \frac{1}{2}$ maximizes precision
  - why might $e(x)$ differ from $\frac{1}{2}$?
- Can you imagine settings where this is not a good design?

Completely Randomized Experiment

- In a completely randomized experiment, sample sizes for each treatment group are fixed in advance
- $N_T =$ size of treatment group
- $N_C =$ size of control group
- Often $N_T = N/2$, but not always
- $e(x) = NT/N$
- Group sizes are the only restriction
Stratified or conditionally randomized experiment

- In a stratified randomized experiment, population is partitioned into blocks or strata within which individuals are similar with respect to one or more covariates
  - Strata may correspond to individuals with similar education, demographics, etc.
  - Strata may correspond to individuals who fall in different risk categories for diabetes
- Individuals are completely randomized within each block/strata
- Ensures balance for important covariate(s)
- Also called blocking
- Heuristic: Block what you can, randomize what you cannot
Stratified or conditionally randomized experiments

- The probability of an individual being treated depends on, say, result of some test that is indicative of criticality (critical versus non-critical)
- Does exchangeability hold?
- No, because the treated and untreated groups may be unbalanced in terms of prognosis
- But such an experiment is simply a sequence of two or more conditionally randomized experiments (say based on the value of \( L \), e.g., critical versus non-critical)
- Hence, we can use standardization or inverse probability weighting to obtain average causal effect by adjusting for \( L \)

\[
CRR = \frac{Pr(Y_a=1)}{Pr(Y_a=0)} = \frac{\sum Pr(Y=1|L=1, A=1)Pr(L=1)}{\sum Pr(Y=1|L=1, A=0)Pr(L=1)}
\]

Paired randomized experiments

- In a paired randomized experiment, individuals are first matched into pairs that are similar with respect to covariates
- Within each pair, randomize which individual is treated
- Special case of blocking
- Goal: improve covariate balance and increase precision
- Also called matched pairs experiments
Summary of Randomized Experiments

- Randomly assigning individuals to treatments
  - Creates balanced treatment groups
  - Eliminates confounding of treatment and outcome by confounders

- Four types of classical randomized experiments:
  - Bernoulli randomized experiment
  - Completely randomized experiment
  - Stratified or conditionally randomized experiment
  - Paired randomized experiment
Sampling from the population

• Suppose we also want to consider random sampling from the population (in addition to random assignment)

• How do things change?

• Replace $\tau = \bar{Y}_{a=1} - \bar{Y}_{a=0}$ with $\tau = E(\bar{Y}_{a=1} - \bar{Y}_{a=0})$

• Define unbiased estimator

• Specify true sample variance

• Apply familiar $t$-test for hypothesis testing

Causal effect estimation
Diet Cola and Calcium

• Does drinking diet cola leach calcium from the body?

• 16 healthy women aged 18-40 were randomly assigned to drink 24 ounces of either diet cola or water

• Their urine was collected for 3 hours, and calcium excreted was measured (in mg)

• Is there a significant difference?

Diet Cola and Calcium

<table>
<thead>
<tr>
<th>Drink</th>
<th>Calcium Excreted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet cola</td>
<td>50</td>
</tr>
<tr>
<td>Diet cola</td>
<td>62</td>
</tr>
<tr>
<td>Diet cola</td>
<td>48</td>
</tr>
<tr>
<td>Diet cola</td>
<td>55</td>
</tr>
<tr>
<td>Diet cola</td>
<td>58</td>
</tr>
<tr>
<td>Diet cola</td>
<td>61</td>
</tr>
<tr>
<td>Diet cola</td>
<td>58</td>
</tr>
<tr>
<td>Diet cola</td>
<td>56</td>
</tr>
<tr>
<td>Water</td>
<td>48</td>
</tr>
<tr>
<td>Water</td>
<td>46</td>
</tr>
<tr>
<td>Water</td>
<td>54</td>
</tr>
<tr>
<td>Water</td>
<td>45</td>
</tr>
<tr>
<td>Water</td>
<td>53</td>
</tr>
<tr>
<td>Water</td>
<td>46</td>
</tr>
<tr>
<td>Water</td>
<td>53</td>
</tr>
<tr>
<td>Water</td>
<td>48</td>
</tr>
</tbody>
</table>
Test Statistic

- A test statistic, $T$, can be any function of:
  - the observed outcomes, $Y_{obs}$
  - the treatment assignments, $W$
  - the covariates, $X$

- The test statistic must be a scalar (one number)

- Examples:
  - Difference in means
  - Regression coefficients
  - Rank statistics

Diet Cola and Calcium

- Difference in sample means between treatment group (diet cola drinkers) and control group (water drinkers)

$$T_{obs} = \frac{\sum_{i=1}^{N_{\text{Treated}}} w_i Y_i^{A=1}}{N_{\text{Treated}}} - \frac{\sum_{i=1}^{N_{\text{Control}}} (1-w_i) Y_i^{A=0}}{N_{\text{Control}}}$$

$$= 6.875 \, mg$$
Key Question

• Is a difference of 6.875 mg more extreme than we would have observed, just by random chance, if there were no difference between diet cola and water regarding calcium excretion?

• What types of statistics would we see, just by the random assignment to treatment groups?

p-value

• $T$: A random variable

• $T_{obs}$: the observed test statistic computed in the actual experiment

• The p-value is the probability that $T$ is as extreme as $T_{obs}$, if the null is true

• GOAL: Compare $T_{obs}$ to the distribution of $T$ under the null hypothesis, to see how extreme $T_{obs}$ is

• SO: Need distribution of $T_{obs}$ under the null
Randomness

- In Fisher’s framework, the only randomness is the treatment assignment: \( W \)
- The potential outcomes are considered fixed
- The distribution of \( T \) arises from the different possibilities for \( W \)
- For a completely randomized experiment, \( N \) choose \( N_T \) possibilities for \( W \)

Sharp Null Hypothesis

- Fisher’s sharp null hypothesis is that there is no treatment effect:

\[ H_0: Y_i^{A=1} = Y_i^{A=0} \text{ for all } i \]

- Note: this null is stronger than the typical hypothesis of equality of the means
- Advantage of Fisher’s sharp null: under the null, all potential outcomes “known”!
  - Why?
  - Because \( Y_i^{A=1} = Y_i^{A=0} \) for all \( i \)
  - At least one of the two outcomes is observed for all \( i \)
Diet Cola and Calcium

• There is NO EFFECT of drinking diet cola (as compared to water) regarding calcium excretion

• So, for each person in the study, their amount of calcium excreted would be the same, whether they drank diet cola or water

Sharp Null Hypothesis

• Key point: under the sharp null, the vector $Y_{\text{obs}}$ does not change with different $W$

• Therefore we can compute $T$ exactly under the null for each different $W$!

• Assignment mechanism completely determines the distribution of $T$ under the null

• (Why is this not true without sharp null?)
Randomization Distribution

- The **randomization distribution** is the distribution of the test statistic, $T$, assuming the null is true, over all possible assignment vectors, $W$

- For each possible assignment vector, compute $T$ (keeping $Y_{obs}$ fixed, because we are assuming the null)

- The randomization distribution gives us *exactly* the distribution of $T$, assuming the sharp null hypothesis is true

Diet Cola and Calcium

- 16 choose 8 = 12,870 different possible assignment vectors

- For each of these, calculate $T$, the difference in sample means, keeping the values for calcium excretion fixed
Diet Cola and Calcium

![Exact Randomization Distribution]

**Exact p-value**

- From the randomization distribution, computing the p-value is straightforward:

- The *exact p-value* is the proportion of test statistics in the randomization distribution that are as extreme as $T_{obs}$

- This is exact because there are no distributional assumptions – we are using the exact distribution of $T$
Diet Cola and Calcium

- If there were no difference between diet cola and water regarding calcium excretion, only 5/1000 of all randomizations would lead to a difference as extreme as 6.875 mg (the observed difference)

- Drinking diet cola probably does rob your body of calcium!
Notes

- This approach is completely nonparametric – no model specified in terms of a set of unknown parameters
- We don’t model the distribution of potential outcomes (they are considered fixed)
- No modeling assumptions or assumptions about the distribution of the potential outcomes

Approximate p-value

- For larger samples, the number of possible assignment vectors (N choose N_t) gets very large
- Enumerating every possible assignment vector becomes computationally difficult
- It’s often easier to simulate many (10,000? 100,000?) random assignments
Approximate p-value

• Repeatedly randomize units to treatments, and calculate test statistic keeping $Y_{\text{obs}}$ fixed

• If the number of simulations is large enough, this randomization distribution will look very much like the exact distribution of $T$

• Note: estimated p-values will differ slightly from simulation to simulation. This is okay!

• The more simulations, the closer this approximate $p$-value will be to the exact $p$-value
Example: Sleep or Caffeine?

- Is sleep or caffeine better for memory?
- 24 adults were given a list of words to memorize, then randomly divided into two groups
- During a break one group took a nap for an hour and a half, while the other group stayed awake and then took a caffeine pill after an hour
- Y: number of words recalled

Mednick et al., "Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory", Behavioral Brain Research, 2008; 193: 79-86.
Fisher versus Neyman

• At the same time Fisher was developing his framework for inference, Neyman was developing his own framework...

• Fisher: more focused on testing
  • is there a causal effect?
  • p-values

• Neyman: more focused on estimation
  • average treatment effect
  • unbiased estimators
  • confidence intervals

Neyman’s Plan for Inference

• Define the **estimand**
• Look for an **unbiased estimator** of the **estimand**
• Calculate the **true sampling variance** of the **estimator**
• Look for an **unbiased estimator** of the **true sampling variance** of the **estimator**
• Assume approximate normality to obtain p-value and confidence interval
Finite Sample versus Super Population

- Finite sample inference:
  - Only concerned with units in the sample
  - Only source of randomness is random assignment to treatment groups

- Super population inference:
  - Extend inferences to greater population
  - Two sources of randomness: random sampling, random assignment
  - “repeated sampling”

Sleep or Caffeine

- Fisher: Is there any difference between napping or staying awake and consuming caffeine, regarding number of words recalled?

- Neyman: On average, how many more words are recalled if a person naps rather than stays awake and consumes caffeine?
Estimand: Average causal effect

- In the finite sample setting, the average causal effect of treatment is defined as:

\[
\tau = \bar{Y}_{a=1} - \bar{Y}_{a=0} = \frac{\sum_{i=1}^{N} Y_{i}^{a=1}}{N} - \frac{\sum_{i=1}^{N} Y_{i}^{a=0}}{N}
\]

Estimator

- For completely randomized experiments, \( \hat{\tau} \) is an unbiased estimator of \( \tau \)

\[
\hat{\tau} = \bar{Y}_{A=1} - \bar{Y}_{A=0} = \frac{\sum_{i=1}^{N} w_{i} Y_{i}^{A=1}}{N_{Treated}} - \frac{\sum_{i=1}^{N} (1 - w_{i}) Y_{i}^{A=0}}{N_{Control}}
\]

where \( w_{i} = 1 \) if the \( i \)th individual is treated and 0 otherwise.
Sleep versus Caffeine

- Estimand: the average word recall for all 24 people if they had slept – average word recall for all 24 people if they had caffeine
- Hence, the estimator assumes exchangeability of the treated and untreated populations
- Estimate varies from one random assignment to another

Unbiased

- An estimator is unbiased if the average of the estimator computed over all assignment vectors (W) will equal the estimand
- The estimator is unbiased if

$$E(\hat{\tau}) = \tau$$
Sleep versus Caffeine

• Estimator is unbiased if $E(\hat{\tau}) = \tau$
• For completely randomized experiments,

$$\hat{\tau} = \frac{\sum_{i=1}^{N} w_i Y_i^{A=1}}{N_{Treated}} - \frac{\sum_{i=1}^{N} (1 - w_i) Y_i^{A=0}}{N_{Control}}$$

is an unbiased estimator of

$$\tau = \frac{\sum_{i=1}^{N} Y_i^{a=1}}{N} - \frac{\sum_{i=1}^{N} Y_i^{a=0}}{N}$$

if the treated and untreated populations are exchangeable

---

Sleep vs Caffeine

• Estimand: the average word recall for all 24 people if they had napped – average word recall for all 24 people if they had caffeine

• Estimator:

$$\hat{\tau} \equiv \bar{Y}_S^{obs} - \bar{Y}_C^{obs} = 15.25 - 12.25 = 3$$
Neyman’s Plan for Inference

- Define the estimand
- Look for an unbiased estimator of the estimand
- Calculate the true sampling variance of the estimator
- Look for an unbiased estimator of the true sampling variance of the estimator
- Assume approximate normality to obtain p-value and confidence interval

True Variance over assignments

\[
Var(Y_{A=1} - Y_{A=0}) = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C} - \frac{S_{TC}^2}{N}
\]

\[
S_T^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (Y_i^{A=1} - \bar{Y}_{A=1})^2
\]

\[
S_C^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (Y_i^{A=0} - \bar{Y}_{A=0})^2
\]

\[
S_{TC}^2 = \frac{1}{N - 1} \sum_{i=1}^{N} ((Y_i^{A=1} - Y_i^{A=0}) - (\bar{Y}_{A=1} - \bar{Y}_{A=0}))^2
\]

Sample variance of potential outcomes under treatment and control, for all units.

Derivation omitted
Note

\[ S_{TC}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left( Y_i^{A=1} - Y_i^{A=0} - (Y_i^{A=1} - Y_i^{A=0}) \right)^2 \]

- Always positive
- Equal to zero if the treatment effect is constant for all individuals
- Related to the correlation between the observed treatment outcome and control outcome (perfectly correlated if the treatment effect is constant)

Neyman’s Plan for Inference

- Define the estimand
- Look for an unbiased estimator of the estimand
- Calculate the true sampling variance of the estimator
- Look for an unbiased estimator of the true sampling variance of the estimator
  - Impossible!
- Assume approximate normality to obtain p-value and confidence interval
Estimator of Variance (of estimator of $\tau$)

\[
\widehat{\text{Var}}(\hat{\tau}) = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C}
\]

\[
S_T^2 = \frac{1}{N_T - 1} \sum_{i=1}^{N} W_i (Y_i^{A=1} - \bar{Y}^{A=1})^2
\]

Sample variances of observed outcomes under treatment and control

\[
S_C^2 = \frac{1}{N_C - 1} \sum_{i=1}^{N} (1 - W_i) (Y_i^{A=0} - \bar{Y}^{A=0})^2
\]

Estimator of Variance

- This is the standard variance estimate used in the familiar $t$-test
- For finite samples, this is may be an overestimate of the true variance
- Resulting inferences may be too conservative (confidence intervals too wide, $p$-values too large)
Sleep vs Caffeine

\[ \text{var}(\hat{\tau}) = \frac{s_T^2}{N_T} + \frac{s_C^2}{N_C} \]

\[ = \frac{3.3^2}{12} + \frac{3.5^2}{12} \]

\[ = 1.958 \]

Central Limit Theorem

- Neyman’s inference relies on the central limit theorem: sample sizes must be large enough for the distribution of the estimator to be approximately normal
- Depends on the sample size AND the distribution of the outcome
- Need larger \( N \) if the distribution is highly skewed, or some individuals are outliers or if some outcomes are rare
Confidence Intervals

\[ \hat{t} \pm z^* \sqrt{\text{var}(\hat{t})} \]

- \( z^* \) (or \( t^* \)) is the value leaving the desired percentage in between \(-z^*\) and \(z^*\) in the standard normal distribution
- Confidence intervals due to Neyman!
- In the finite sample case, confidence interval may be too wide, and hence inference too conservative
- A 95% confidence interval will contain the estimand at least 95% of the time

Hypothesis Testing

- Fisher \( H_0: Y_i^{A=1} = Y_i^{A=0} \) for all \( i \) (Sharp null hypothesis)
- Neyman: null hypothesis of no treatment effect on average
  \[ H_0: \overline{Y_{a=1}} - \overline{Y_{a=0}} = 0 \]
Hypothesis Testing

- Fisher: compare any test statistic to empirical randomization distribution

- Neyman: compare t-statistic to normal or t distribution (relies on large n)

\[ t = \frac{\hat{\tau}}{\text{var}(\hat{\tau})} = \frac{Y_{T}^{\text{obs}} - Y_{C}^{\text{obs}}}{\sqrt{\frac{s_{T}^2}{N_{T}} + \frac{s_{C}^2}{N_{C}}}} \]

(Neyman’s approach is the familiar t-test)

Sleep vs Caffeine

\[ t = \frac{Y_{S}^{\text{obs}} - Y_{C}^{\text{obs}}}{\sqrt{\frac{s_{S}^2}{N_{S}} + \frac{s_{C}^2}{N_{C}}}} = \frac{15.25 - 12.25}{\sqrt{\frac{3.3^2}{12} + \frac{3.5^2}{12}}} = 2.14 \]
Sleep vs Caffeine

Exact Randomization Distribution

Exact p-value = 0.0252

Difference in Means

Super Population

- Suppose we also want to consider random sampling from the population (in addition to random assignment)

- How do things change?
Neyman Inference (Super Population)

1. Define the estimand: $Y_{a=1} - Y_{a=0}$

2. unbiased estimator of the estimand: $\hat{\tau} = \bar{Y}_{T}^{obs} - \bar{Y}_{C}^{obs}$

3. true sampling variance of the estimator $\text{var}(\hat{\tau}) = \frac{\sigma_{T}^{2}}{N_{T}} + \frac{\sigma_{C}^{2}}{N_{C}}$

4. unbiased estimator of the true sampling variance of the estimator $\text{var}(\hat{\tau}) = \frac{\hat{\sigma}_{T}^{2}}{N_{T}} + \frac{\hat{\sigma}_{C}^{2}}{N_{C}}$

5. Assume approximate normality to obtain p-value and confidence interval

Super Population

- Neyman’s approach (and therefore all the familiar t-based inference you are used to) are considering both random sampling from the population and random assignment
### Fisher vs Neyman

<table>
<thead>
<tr>
<th>Fisher</th>
<th>Neyman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal: testing</td>
<td>Goal: estimation</td>
</tr>
<tr>
<td>Considers only random assignment</td>
<td>Considers random assignment and random sampling</td>
</tr>
<tr>
<td>$H_0$: no treatment effect</td>
<td>$H_0$: average treatment effect = 0</td>
</tr>
<tr>
<td>Works for any test statistic</td>
<td>Difference in means</td>
</tr>
<tr>
<td>Exact distribution</td>
<td>Approximate, relies on large $N$</td>
</tr>
<tr>
<td>Works for any known assignment mechanism</td>
<td>Only derived for common designs</td>
</tr>
</tbody>
</table>

### Announcements

- **Take home exam** – to be posted by this Thursday, due next Thursday
- **Term project** – timeline and further guidance to be posted by this Thursday, with proposals due in 2 weeks
  - Please feel free to discuss your project proposal ideas with me in advance during next class period or during the online office hours
Covariates and Covariate Balance

• Pre-treatment variables
• $X: n \times k$ covariate matrix ( $n$ samples, $k$ covariates)
• Balance covariates between treatment groups
• Covariate Balancing
  • Why?
    • Ensure exchangeability of the treated and untreated populations, or more generally, different treatment groups
  • How?
Covariates

- Randomization *should* balance all covariates (observed and unobserved) on average...
- ... but covariates may be imbalanced by random chance, and sometimes better balance is desired

Covariate Balance

- Why is covariate balance important?
- Better covariate balance...
  - provides more meaningful estimates of the causal effect
  - increases precision (reduces variance) of estimator, if covariates are correlated with outcome
  - outcome less variable for similar values of covariates
Covariate Balance

Two options:
- Option 1: force better balance on important covariates by design
- Option 2: correct imbalance in covariates by analysis
Stratified Experiments

- Units are stratified (grouped, blocked) according to covariate(s)
- Subdivide sample into \( J \) homogeneous strata (blocks)
- Randomize units to treatment groups within each stratum
- Often used with important categorical covariates

Bee Stings

- If you are stung by a bee, does that make you more likely to get stung again?
- Scientists dangled 16 muslin-wrapped cotton balls over a beehive, where half of the balls had been previously stung and the other half were fresh
- Outcome: total number of new stingers
- This was repeated for a total of nine trials.

Bee Stings

- Scientists expect the number of stings to vary by trial (different number of bees in the hive, different times of day, different weather, etc.)
  - Each trial can be seen as a different stratum
  - \( J = 9 \) strata
Stratified Experiment

- What to use for a test statistic?

\[ \bar{Y}_T^{obs}(j) = \text{average observed } Y \text{ for treated units in the } j^{th} \text{ strata} \]
\[ \bar{Y}_C^{obs}(j) = \text{average observed } Y \text{ for control units in the } j^{th} \text{ strata} \]

For each strata: \[ \bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j) \]

- How to combine?

\[ T = \sum_{j=1}^{J} \lambda_j (\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j)) \]
Stratified Experiment

\[ T = \sum_{j=1}^{J} \lambda_j \left( \bar{Y}_{T}^{\text{obs}} (j) - \bar{Y}_{C}^{\text{obs}} (j) \right) \]

- What to use for the weights?
- Weights must sum to 1
- Weight by the sample size of each strata, \( N(j) \):

\[ T = \sum_{j=1}^{J} \frac{N(j)}{N} \left( \bar{Y}_{T}^{\text{obs}} (j) - \bar{Y}_{C}^{\text{obs}} (j) \right) \]

Bee Stings

- In this example, the trials are all the same sample size, so just a simple average of the treatment effects for each trial:

\[ T^{\text{obs}} = 12 \]
Inference

- Fisher randomization test
- Easier to simulate randomizations, rather than enumerate all possible allocations
- For each simulated randomization, just randomize within strata
- For randomization test, just simulate randomization scheme actually used in experiment

Bee Stings

Randomization Distributions

- p-value = 0.04
Stratified Experiment

• What to use for an estimator (Neyman)?

• In general... 
  \[ \hat{\tau} = \sum_{j=1}^{J} \lambda_j \hat{\tau}_j \]

• One common option: 
  \( \hat{\tau}_j \) is the estimate within strata \( j \)
  \( \lambda_j \) is the weight given to strata \( j \)

\[ \hat{\tau} = \sum_{j=1}^{J} \frac{N(j)}{N} \left( \bar{Y}_{T}^{\text{obs}}(j) - \bar{Y}_{C}^{\text{obs}}(j) \right) \]

Stratified Experiments

\[ \text{var}(\hat{\tau}) = \sum_{j=1}^{J} \lambda_j^2 \text{var}(\hat{\tau}_j) \]

\[
\text{var}\left( \sum_{j=1}^{J} \frac{N(j)}{N} \left( \bar{Y}_{T}^{\text{obs}}(j) - \bar{Y}_{C}^{\text{obs}}(j) \right) \right) \\
= \sum_{j=1}^{J} \frac{N(j)^2}{N^2} \left( \frac{S_{T,j}^2}{N_T(j)} + \frac{S_{C,j}^2}{N_C(j)} \right) 
\]
Stratified Experiments

- In the case of randomized experiments, stratification can do no harm, and may help
  - Not so in the case of observational data
  - Why?
- Can stratify on more than one covariate
- Strata can be any size
- Stratify what you can; randomize what you cannot

Paired Experiments

- Units are matched into pairs (based on covariate(s))
- Special case of stratified randomized experiment with $N_j = 2$ for each stratum
- Useful when expect difference in potential outcomes within a pair to be much smaller than differences across pairs
Wetsuit Advantage?

• The 2008 Olympics were full of controversy about whether the new wetsuits provide an unfair advantage
• Can a wetsuit really make someone swim faster? How much faster?

Wetsuit Advantage

• Twelve competitive swimmers and triathletes swam 1500m at maximum speed twice each, once wearing a wetsuit and once wearing a regular suit
• Maximum velocity (m/sec) recorded (one of several possible outcomes)
• The order of the trials was randomized
• Each person is one “pair”

Wetsuit Advantage?

- Box plot showing the maximum velocity for Regular Suit and Wetsuit.
- Scatter plot comparing maximum velocity with and without Wetsuit.

93

94
Paired Experiment

- Test statistic / estimate: average of differences across all pairs
- Randomization test: randomize sign of each difference
- Neyman inference: analyze differences as a single variable

Paired Experiments

- Analysis is easy!
- When variability within pairs is much smaller than variability across pairs, can get huge gains in precision
- Better precision translates into higher power for tests (lower p-values) and narrower confidence intervals
Rerandomizing Randomized Experiments

- Randomized experiments are the “gold standard” for estimating causal effects
- Why?
  - They yield unbiased estimates
  - They balance covariates across treatment groups
Covariate Balance - Gender

- What if you get a “bad” randomization?
- What would you do?
- Can you rerandomize? When? How?

Covariate Imbalance

- The more covariates we have, the more likely it is that at least one covariate will be imbalanced across treatment groups
- With just 10 independent covariates, the probability of a significant difference ($\alpha = .05$) for at least one covariate is $1 - (1 - .05)^{10} = 40%$
- Covariate imbalance is not limited to rare “unlucky” randomizations
Randomized Experiments – The fine print

• Randomized experiments are the “gold standard” for estimating causal effects
• Why?
  • They yield unbiased estimates
  • They eliminate confounding factors ... on average!
• For any particular experiment, covariate imbalance is possible (and likely!), and conditional bias exists

1) Collect covariate data
(Re)randomize subjects to treated and control

2) Specify criteria for determining when a randomization is unacceptable -- based on covariate balance
Check covariate balance

unacceptable

acceptable

3) Conduct experiment

4) Analyze results (with a randomization test)
### Rerandomization Criterion

- Let $\mathbf{x}$ be the covariate matrix
- Let $\mathbf{W}$ be the vector of treatment assignments

$$W_i = \begin{cases} 
1 & \text{if } i^{th} \text{ subject is treated} \\
0 & \text{if } i^{th} \text{ subject is control}
\end{cases}$$

- The criterion determining whether a randomization, $\mathbf{W}$, is acceptable should be some function of $\mathbf{x}$ and $\mathbf{W}$

### Potential Outcomes

- Let $y_i(W_i)$ denote the $i^{th}$ unit’s potential outcome under treatment group $W_i$

$$
\tau = \frac{\sum_{i=1}^{n} y_i(1) - \sum_{i=1}^{n} y_i(0)}{n} = \frac{\sum_{i=1}^{n} W_i y_i(1) - \sum_{i=1}^{n} (1-W_i)y_i(0)}{\sum_{i=1}^{n} W_i - \sum_{i=1}^{n} (1-W_i)}
$$
Unbiased estimates

If the treated and control groups are the same size, and if the criteria for rerandomization treats the treated and control groups equally, then

\[ E(\hat{\tau}) = \tau \]

Intuition: For every randomization that is thrown away, there is an exact opposite randomization, \( 1 - W \), that is also thrown away.

Unbiased estimates

If treated and control groups are not the same size, then rerandomization may not yield an unbiased estimate.

Example: Suppose you have one covariate \( x \), and the criteria for rerandomization is \( |\bar{X}_T - \bar{X}_C| < 1 \)

Units with more extreme \( x \) values will be more likely to be in the larger treatment group.
Rerandomization Test

- **Randomization Test:**
  - Simulate randomizations to see what the statistic would look like just by random chance, if the null hypothesis were true

- **Rerandomization Test:**
  - A randomization test, but for each simulated randomization, follow the same rerandomization criteria used in the experiment

  As long as the simulated randomizations are done using the same randomization scheme used in the experiment, this will give accurate p-values

Alternatives for Analysis

- **t-test:**
  - Too conservative
  - Significant results can be trusted

- **Regression:**
  - Regression including the covariates that were balanced on using rerandomization more accurately estimates the true precision of the estimated treatment effect
  - Assumptions are less dangerous after rerandomization because groups should be well balanced
• If you have more than one covariate that you think may be associated with the outcome, what would you use as criteria for an acceptable randomization?
  • The obvious choice may be to set limits on the acceptable balance for each covariate individually
  • This destroys the joint distribution of the covariates

We use Mahalanobis Distance, $M$, to represent multivariate distance between group means:

$$M = (\bar{X}_T - \bar{X}_C)'\text{cov}(\bar{X}_T - \bar{X}_C)^{-1}(\bar{X}_T - \bar{X}_C)$$

$$= \left(\frac{1}{n_T} + \frac{1}{n_C}\right)^{-1}(\bar{X}_T - \bar{X}_C)'\text{cov}(X)^{-1}(\bar{X}_T - \bar{X}_C)$$

Choose $a$ and rerandomize when $M > a$
Choice of $p_a$

- Choice of acceptance probability entails a tradeoff between a desire for better balance and computational cost.
- The number of randomizations needed to get an acceptable one is $\text{Geometric}(p_a)$, so the expected number needed is $1/p_a$. 
Rerandomization Based on $M$  

- Since $M$ follows a known distribution, easy to specify the proportion of accepted randomizations  
- $M$ is affinely invariant (unaffected by affine transformations of the covariates)  
- Correlations between covariates are maintained  
- The balance improvement for each covariate is the same and also for any linear combination of the covariates  

Rerandomization  

- Rerandomization improves covariate balance between the treatment groups (and hence makes the treatment groups “more exchangeable”)  
- If the covariates are correlated with the outcome, rerandomization also increases precision in estimating the treatment effect, thus increasing the ability to detect a significant causal effect using fewer subjects