Randomized Experiment

- Identify causal effects

Average Treatment Effect on the Treated (ATT) and Selection Bias

Table 2.1

	A	Y	Y^0	Y^1
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaestus	1	1	?	1
Aphrodite	1	1	?	1
Cyclope	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

- Fundamental problem of causal inference: Only one of the two counterfactual outcomes is known for each individual, that is, the one corresponding to the treatment level that was actually received.

- Good to compute associational measures $E(Y|A = 1) - E(Y|A = 0)$.

Observed data and potential outcomes

$E(Y|A=1) - E(Y|A=0)$

$= E(Y^1|A=1) - E(Y^0|A=0)$ by consistency

$= E(Y^1|A=1) - E(Y^0|A=1) + E(Y^0|A=1) - E(Y^0|A=1)$

$= \underline{E(Y^1 - Y^0|A=1)} + \underline{E(Y^0|A=1) - E(Y^0|A=0)}$

= ATT + Selection Bias,

where

- ATT: Average Treatment Effect on the Treated (ATT)
- Selection Bias: how different the treated and untreated group are in terms of their potential outcome if untreated

How can we eliminate the bias term?

Randomization in Experiments: A study where assignment to treatment is controlled by the researcher

A **randomized** experiment satisfies:

For a=0 or 1,

1. **Positivity**: assignment is probabilistic
   a. $0 < p_i = P[A_i = a] < 1$
   b. No deterministic assignment
2. **Exchangeability**: $P[A_i = a] = P[A_i = a|Y^0, Y^1]$
   Treatment assignment does not depend on any potential outcomes, written as $A \perp (Y^0, Y^1)$

With Randomization (**Positivity** and **Exchangeability** hold):

- Treatment and control groups are random samples from the population
- Control group mean and treatment group mean are both unbiased for the population mean of the potential outcome, i.e.

$$E(Y^a|A=0) = E(Y^a|A=1) = E(Y^a) \text{ for } a=0 \text{ or } 1$$

- **Randomization eliminates selection bias (SB):**

$$SB = E(Y^0|A=1) - E(Y^0|A=0) = E(Y^0) - E(Y^0) = 0$$

- Does Randomization make ATT=ACE?

ATT=$E(Y^1|A=1) - E(Y^0|A=1) = E(Y^1) - E(Y^0)$ =ACE

ACE is identified in a randomized experiment! ☺

Types of Randomized Experiments

1). Completely randomized experiment

- Randomly select n_t from n experimental units to be treated
- Equal selection probability n_t/n
- Selection is not independent between units $E[A_i A_j] = \frac{n_t}{n}\frac{n_t-1}{n-1}$
- Randomization **balances** both observed or unobserved pre-treatment covariates between the treated and untreated in large samples
- Randomization procedure above implies exchangeability, i.e.,

$$A \perp (Y^0, Y^1)$$

Therefore, $\text{ACE} = E(Y^1 - Y^0) = E(Y^1|A=1) - E(Y^0|A=0) = E(Y|A=1) - E(Y|A=0)$ → ACE is identified, estimated by

$$\frac{1}{n_t}\sum_{i=1}^{n_t} Y_i - \frac{1}{n-n_t}\sum_{i=1}^{n-n_t} Y_i$$

Example – completely randomized experiment:

i	Y	A	Y^1	Y^0
1	2	1	2	?
2	0	1	0	?
3	1	0	?	1
4	3	0	?	3

Estimate average casual effect $E(Y^1) - E(Y^0)$.

Random assignment implies exchangeability that

$E(Y^1|A = 1) = E(Y^1|A = 0) = E(Y^1)$ and

$E(Y^0|A = 1) = E(Y^0|A = 0) = E(Y^0)$

$\tau = \tau_{ATT} = E(Y^1|A = 1) - E(Y^0|A = 0) = E(Y|A = 1) - E(Y|A = 0)$

estimated by $\hat{\tau} = \frac{2+0}{2} - \frac{1+3}{2} = -1$

2). Stratified/Conditional randomized experiment

- Form blocks based on covariates L
- Within blocks, completely randomized experiment
- Conditional exchangeability: $A \perp (Y^0, Y^1)|L$

3). Pair randomized experiments:

- Stratified randomized experiment where each block has 2 units
- One unit in each pair receives treatment
- Extreme version
- Matched pair design

Goal: Remove 'bad randomization' where **covariates** are related to treatment assignment by chance.

Identify Causal Effect under Design 2

$E(Y^1 - Y^0)$

$= E_l[E(Y^1 - Y^0|L = l)]$ by **law of total expectation**

$= E_l[E(Y^1|A = 1, L = l) - E(Y^0|A = 0, L = l)]$

by **conditional exchangeability**

$= E_l[E(Y|A = 1, L = l) - E(Y|A = 0, L = l)]$ by **consistency**

Identifiable!

Table 2.2

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Example

Design 1: Randomly selected 65% of the individuals in the population of patients and transplanted a new heart to each of the selected individuals

- Completely randomized experiments
- $Y^a \perp A$, for all a, holds

- Estimate the average causal effect, e.g.,

$$RR = \frac{p(Y^{a=1} = 1)}{p(Y^{a=0} = 1)}$$

  - $p(Y^{a=1} = 1)$*: counterfactual risk of death had every patient in the population been heart transplanted.*
  - $p(Y^{a=0} = 1)$*: counterfactual risk of death had every patient in the population not been heart transplanted.*

$$\widehat{RR} = \frac{\hat{p}(Y^{a=1}=1)}{\hat{p}(Y^{a=0}=1)} = (7/13)/(3/7) = 1.26$$

If all 20 patients have heart transp., the causal risk of mortality is 1.26 times the risk if no one has heart transplant.

Design 2: Classify all individuals as being in either critical (L=1) or noncritical (L=0) condition. Then randomly selected 75% of the individuals in critical condition and 50% of those in noncritical condition, and transplanted a new heart to each of the selected individuals.

- Stratified randomized experiments
- $Y^a \perp A|L$ holds
- Each treatment group (A=0 or 1) has a different proportion of subjects with bad condition (69%=9/13 treated versus 43%=3/7 untreated individuals were in critical condition.)
- This imbalance indicates that the risk of death in the treated, had they remained untreated, would have been higher than the risk of death in the untreated. $Y^a \perp A$ doesn't hold

Causal RD and RR under Design 2

- Compute stratum specific causal effect
- Methods to compute effect measures: Standardization and Inverse Probability Weighting

Method 1: Standardization

- Standardized mean

$$E(Y^a) = p(Y^{A=a} = 1) = \sum_l E(Y|L = l, A = a) \times Pr(L = l)$$

In the presence of conditional exchangeability, this standardized risk can be interpreted as the (counterfactual) risk that would have been observed had all the persons in the population been treated (if A=1) or untreated (if A=0).

- Causal risk ratio

$$\frac{p(Y^{a=1}=1)}{p(Y^{a=0}=1)}$$

can be computed by standardization as

$$\frac{\sum_l p(Y=1|L=l,\boldsymbol{A}=\mathbf{1})\times \Pr(L=l)}{\sum_l p(Y=1|L=l,\boldsymbol{A}=\mathbf{0})\times Pr(L=l)} = \frac{\left(\frac{1}{4}\times\frac{8}{20}+\frac{6}{9}\times\frac{12}{20}\right)}{\left(\frac{1}{4}\times\frac{8}{20}+\frac{2}{3}\times\frac{12}{20}\right)} = 1.$$

Method 2: Inverse probability weighting (IPW)

The data in table 2.2 can be displayed as

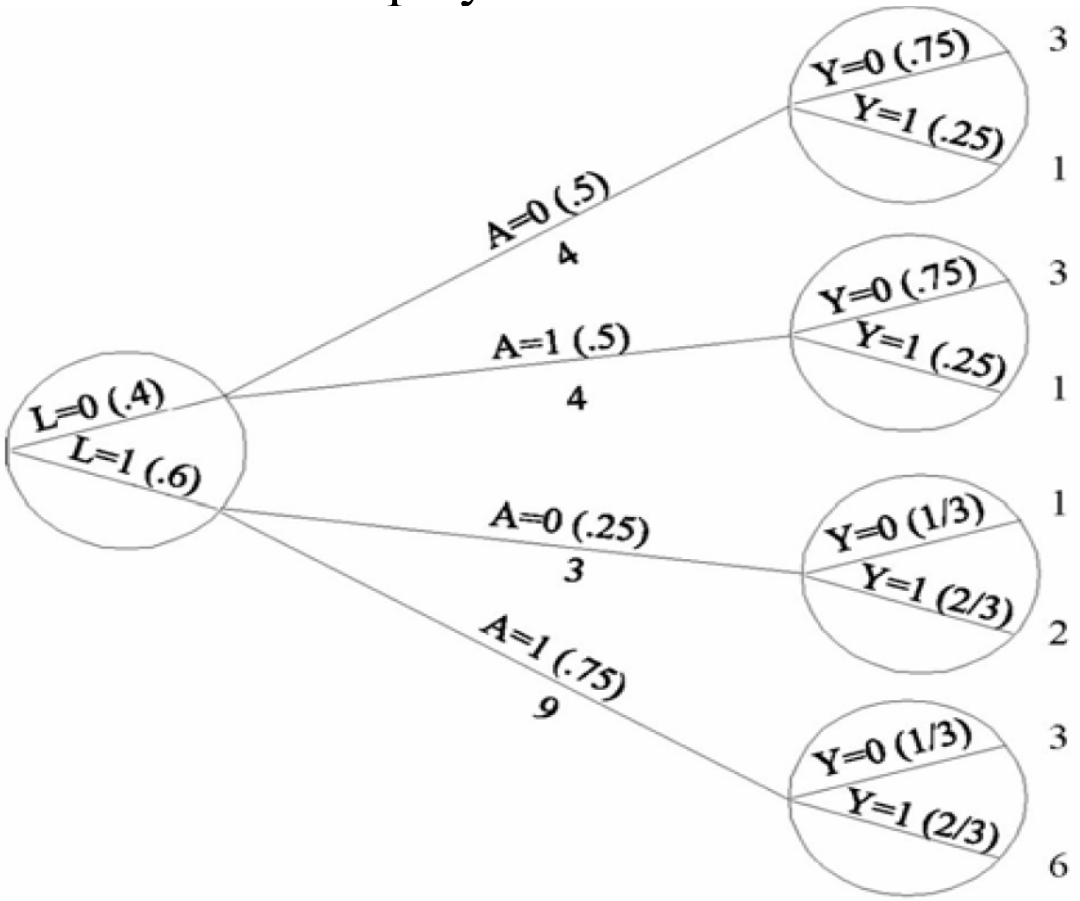


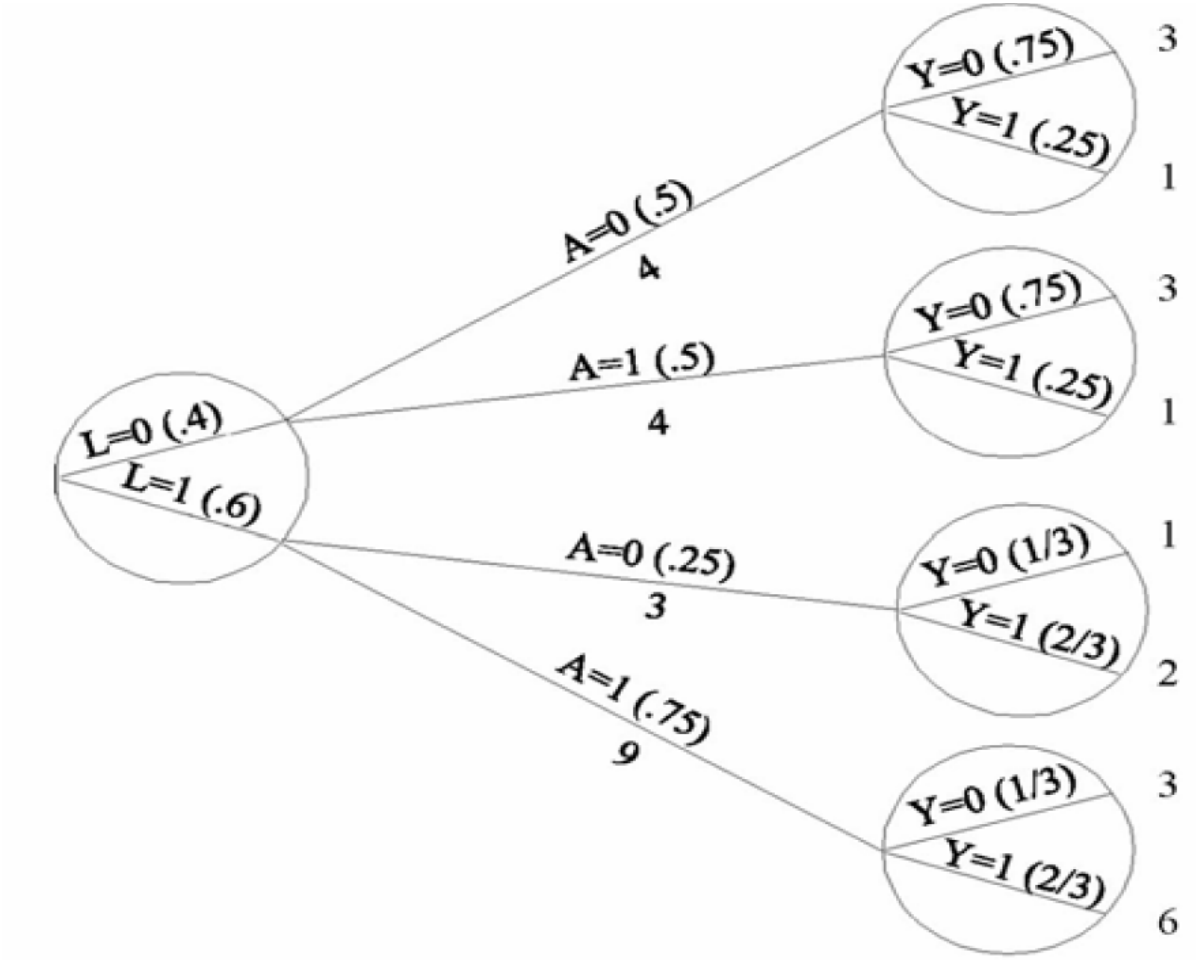
Fig1. A population of 20 patients with prognosis factor L, treatment A and outcome Y

- Use the tree to estimate causal risk ratio $\frac{p(Y^{a=1}=1)}{p(Y^{a=0}=1)}$:
  - $p(Y^{a=0} = 1)$: the counterfactual risk of death had everybody in the population remained untreated:
    - 4 of 8 persons with L=0 is untreated and 1 of 4 died (=1/4). How many deaths would have occurred had all the 8 persons with L=0 remained untreated? 8×1/4=2
    - 3 of 12 persons with L=1 is untreated and 2 of 3 died (=2/3). How many deaths would have occurred had all the 12 persons with L=1 remained untreated? 12×2/3=8

$$\hat{p}(Y^{a=0} = 1) = \frac{2 + 8}{20} = 0.5$$

  - $p(Y^{a=1} = 1)$: the counterfactual risk of death had everybody in the population remained treated: $\hat{p}(Y^{a=1} = 1) = \frac{2+8}{20} = 0.5$
  - Therefore, the causal risk ratio $\frac{\hat{p}(Y^{a=1}=1)}{\hat{p}(Y^{a=0}=1)} = 1$ under the assumption of conditional exchangeability.

Basic idea: Create a pseudo-population by weighting each person i in the population by the inverse of probability of being assigned to the observed treatment A_i



w_i

2

2

4

1.33

Define

$$w_i = \frac{1}{p[A_i|L_i]}$$

IPW estimators are

$$\hat{p}(Y^{a=0} = 1) = \frac{1}{n}\sum_i^n I(A_i = 0) w_i Y_i = \frac{1}{20}(1 \times 2 + 2 \times 4) = 0.5$$

$$\hat{p}(Y^{a=1} = 1) = \frac{1}{n}\sum_i^n I(A_i = 1) w_i Y_i = \frac{1}{20}(1 \times 2 + 6 \times 1.33) = 0.5$$

Therefore,

$$\widehat{RR} = \frac{\hat{p}(Y^{a=1} = 1)}{\hat{p}(Y^{a=0} = 1)} = 1$$

In Summary - Randomized experiments

- **Advantages**
  - Provide strongest and most direct evidence for causality
  - Strong internal validity: estimate ACE for our particular sample if randomization really works (no difference between treated and untreated that affect the outcome and can be controlled for)

- **Disadvantages**
  - Ethical issues: not suitable for all research questions
  - Feasibility and Cost
  - Still some risk of bias (threat to internal validity)
    - Failure of randomization (small sample size)
    - Noncompliance with experimental protocol
  - External validity (can we extrapolate our estimates to other population?) Non-representative sample
    - Convenience sample / not from population of interest

For example,

It is questionable that an **ethical** committee would have approved our heart transplant study. Hearts are in short supply and society favors assigning them to subjects who are more likely to benefit from the transplant, rather than assigning them randomly among potential recipients. Also one could question the **feasibility** of the study even if ethical issues were ignored: double-blind assignment is impossible, individuals assigned to medical treatment may not resign themselves to forego a transplant, and there may not be compatible hearts for those assigned to transplant.

Observational Studies

Observational studies

- Advantages
  - Cheaper
  - Larger numbers
  - Longer follow-up
  - Likely to be more generalizable because include more representative sample of population
  - Take place in normal health care settings
  - Efficient use of available data
- Disadvantages
  - Non-randomized allocation to exposure of interest so strong likelihood of bias and confounding
  - Data more likely to be incomplete and of poorer quality

In Randomized Experiments:

Randomization guarantees that potential confounders are statistically independent of treatment group assignment.

- Exchangeability in completely randomized experiment:
  Interpretation: the treated and the untreated are exchangeable because the treated, had they remained untreated, would have experienced the same average outcome as the untreated did, and vice versa.
- Conditional Exchangeability in stratified randomized experiment:
  Interpretation: $Y^a \perp A|L$ holds in because, within levels of L, all other predictors of the outcome are equally distributed (**Balance**) across different treatment groups.

Balance means that a covariate has identical sample distributions across all treatment groups.

Question: which of the following has balanced distribution in smoking status across different treatment groups?

	High dose	**Low dose**
Smokers	30	40
Non-smokers	15	20

	High dose	**Low dose**
Smokers	30	30
Non-smokers	15	30

In Observational studies

Investigators do not intervene in the assignment of hearts but rather they observe which persons happen to receive them. Assume Table 2.2 displays the data collected for this observational study. Investigators **believe** that,

In critical condition patients, had treated patients stayed untreated, they would have had the same mortality risk as patients who actually untreated. Similarly for patients in non-critical condition.

←→ Treated and untreated are exchangeable within levels of L

$$Y^a \perp A|L$$

$E(Y^1 - Y^0|L = l)$ would be identifiable because

$= E(Y^1|L = l, A = 1) - E(Y^0|L = l, A = 0)$ by conditional exchangeability

$= E(Y|L = l, A = 1) - E(Y|L = l, A = 0)$ by consistency

Therefore,

Observational study can be viewed as a conditional randomized experiment (cRE in design 2) in which, different from cRE

- The conditional probabilities of receiving treatment are not chosen by the investigator, but are observed and can be estimated from the data, and
- Conditional exchangeability is not guaranteed but only assumed based on experts' available substantive knowledge.

  There might exist unmeasured confounding, which is out of the control of the researchers.

Bias due to unbalanced confounders

Example1, *L=Smoking status* as *confounder, but unbalanced distribution between treatment groups*

Smoking is an important predictor of the outcome Y, and has different distributions in the treated and the untreated (e.g., smokers have higher probability of receiving treatment).

Suppose an additive model holds,

$$E(Y|A,L) = \alpha_A + \beta_L \text{ for } A \text{ and } L = 0,1$$

where $\alpha_1 - \alpha_0$ is treatment effect and $\beta_1 - \beta_0$ is smoking effect.

- If smoking is controlled for, conditional on smoking status:

  In non-smokers, the treatment effect is

$$E(Y|A=1,L=0)-E(Y|A=0,L=0)=\alpha_1-\alpha_0$$

  In smokers, the treatment effect is

$$E(Y|A=1,L=1)-E(Y|A=0,L=1)=\alpha_1-\alpha_0$$

  Therefore, overall ACE is

$$E(Y|A=1)-E(Y|A=0)=\alpha_1-\alpha_0$$

- What if smoking is unbalanced since we do NOT control for smoking:

  For treated subjects, the mean response is

$$E(Y|A=1) = E(Y|A=1,L=1)p_1 + E(Y|A=1,L=0)(1-p_1)$$
$$= \alpha_1 p_1 + \beta_1 p_1 + \alpha_1(1-p_1) + \beta_0(1-p_1)$$
$$= \alpha_1 + \beta_0 + p_1(\beta_1 - \beta_0),$$

  For untreated subjects, the mean response is

$$E(Y|A=0) = E(Y|A=0,L=1)p_0 + E(Y|A=0,L=0)(1-p_0)$$
$$= \alpha_0 p_0 + \beta_1 p_0 + \alpha_0(1-p_0) + \beta_0(1-p_0)$$
$$= \alpha_0 + \beta_0 + p_0(\beta_1 - \beta_0),$$

  where $p_1 = p(L=1|A=1)$ and $p_0 = p(L=1|A=0)$ are the proportion of smokers in the treated and control group.

Therefore,

The treatment effect without adjusting for the confounder, smoking status, is

$$E(Y|A=1) - E(Y|A=0) = \alpha_1 - \alpha_0 + (p_1 - p_0)(\beta_1 - \beta_0).$$

If smoking rates are not the same in the treated and in the untreated, the estimates without controlling for smoking status are biased by $(p_1 - p_0)(\beta_1 - \beta_0)$.

Example2 - Perfect Doctor

- Suppose that a doctor prescribes surgery (A=1) or drug (A=0) for a certain condition
- Y is years of post-treatment survival, in a pretend world:

Unit	Y^0	Y^1	$Y^1 - Y^0$
1	1	7	6
2	6	5	-1
3	1	5	4
4	8	7	-1
Average	4	6	**2**

If all patients receive surgery, the average years of survival is about 2 years longer than that if all patients take drugs.

- In real world, the doctor knows enough about the potential outcomes of the patients so assigns each patient the treatment that is more beneficial to that patient. Patients 1 and 3 will receive surgery and patients 2 and 4 will receive drug treatment.

Unit	A	Y
1	1	7
2	0	6
3	1	5
4	0	8
Average		6.5

Observed difference in means between groups of A=1 and A=0 is 6-7=-1, that is, the average survival year for the surgery group is 1 year less than that for the drug group →

Misleading when we have effectively pretended we had an *ignorable* treatment assignment when in fact we did not.

Therefore,

- The assignment mechanism depended on the potential outcomes and was nonignorable
- We can reach invalid conclusions if we look at the observed values of potential outcomes without considering how the treatments were assigned.

A major challenge in causal inference from observational studies is finding a statistic that is appropriate for the actual assignment mechanism!

Methods

- Standardization or IPW weighting
  - Both methods assume *conditional exchangeability*
  - Covered under design 2 in the randomized experiments
  - Both methods yielded the same results in simple settings, but slightly different in practical applications because they are based on different modeling assumptions.

Example: Y: test score; A: program enrollment; L: parents' income

Estimate causal effect from A to Y

1. *Standardization* method

   Step1: Cut the income into 20% slices based on the marginal income distribution

   Step2: Estimate treatment effect for the fine slices separately

$$\hat{\tau}_i = \bar{Y}_i(A = 1) - \bar{Y}_i(A = 0) \text{ for } i = 1, \dots, 5$$

Step3: Form a weighted average

$$\hat{\tau} = w_1\hat{\tau}_1 + w_2\hat{\tau}_2 + \cdots + w_5\hat{\tau}_5,$$

where $w_1 + \cdots + w_5 = 1$ are weights.

This weighted average estimates the overall treatment effect, with confounding effect of parents' income mostly removed.

Exercise: Show $w_i = \frac{V_i^{-1}}{\sum_j V_j^{-1}}$ minimizes the $Var(\hat{\tau})$, where $V_i = Var(\hat{\tau}_i)$.

$Var(\hat{\tau}) = \sum_j^5 w_j^2 V_j$, taking derivative to w_i and $\sum_j w_j = 1$

$2w_iV_i - 2w_5V_5 = 0$, solving for w_i

$w_i = \frac{w_5V_5}{V_i}$ for $i = 1, \ldots, 4$ (eq1)

Using $w_1 + \cdots + w_5 = 1 \rightarrow 1 = w_5V_5\left(\sum_j V_j^{-1}\right)$

$\rightarrow w_5V_5 = \left(\sum_j V_j^{-1}\right)^{-1}$, substituted to (eq1) yields $w_i = \frac{V_i^{-1}}{\sum_j V_j^{-1}}$.

2. *IPW* method

Y: test score; A: program enrollment; L: parents' income

Step1: Estimate the conditional probability of observed program enrollment given the parents' income $p[A_i|L_i]$ and the weight for unit i is $1/p[A_i|L_i]$ for $i = 1, \dots, n$

Step2: Estimate the potential outcome given all students enroll the program $\hat{E}[Y^1] = \frac{1}{n}\sum_{i=1}^{n} \frac{I[A_i=1]}{p[A_i|L_i]} Y_i$

Step3: Estimate the potential outcome given all students do NOT enroll the program $\hat{E}[Y^0] = \frac{1}{n}\sum_{i=1}^{n} \frac{I[A_i=0]}{p[A_i|L_i]} Y_i$

Step4: Estimate the treatment effect

$$\hat{\tau} = \hat{E}[Y^1] - \hat{E}[Y^0]$$

Tools (cont.):

- **Matching**
- **Propensity Score**
- **Marginal Structural Models**

Matching

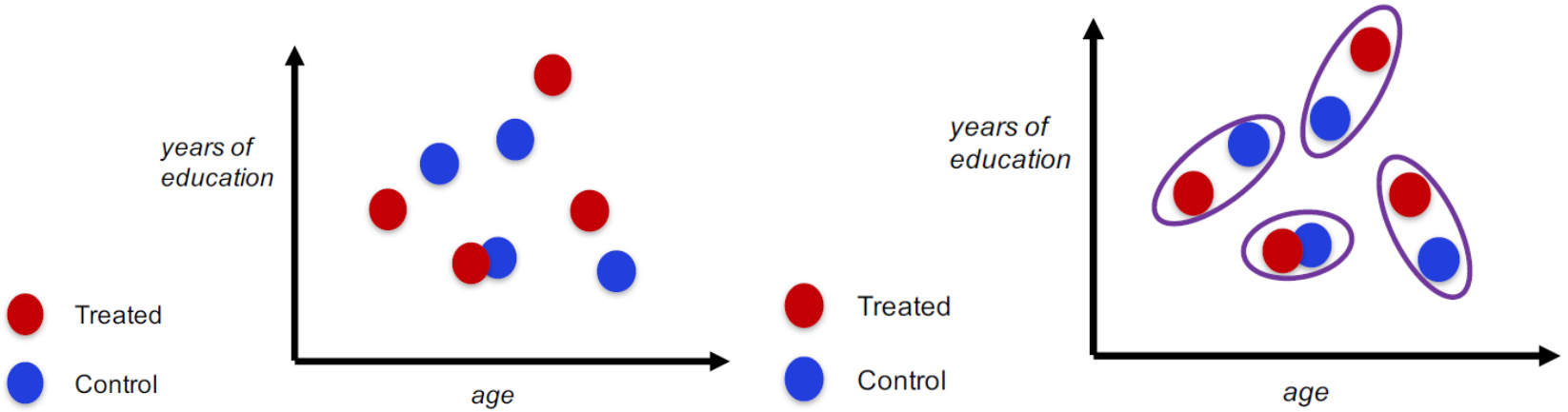
Implementation:

Step 1: Match to find each treated unit's counterfactual identical twin

- **Distance metric**
- **Matching method**

Step 2: Check up on the outcome of matched cases

Step 3: Estimate both ACE and ICE



Distance Metric $d(.,.)$

1. *Exact distance*

$$d(i,j) = \begin{cases} 0 & if\ x_i = x_j \\ \bowtie & if\ x_i \neq x_j \end{cases},$$

2. *Euclidean distance*

   Sqrt of sum of the normalized distances for each covariate k

$$d(i,j) = \sqrt{\sum_{k=1}^{K} \frac{\left(x_{ik} - x_{jk}\right)^2}{\hat{\sigma}_k^2}},$$

   where $\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N}(x_{ik}-\bar{x}_k)^2}{N-1}$ is the sample variance of the k^{th} variable in $x=(x_1,\ldots,x_K)$.

3. *Mahalanobis distance*

- Intuition: If x_{ik} and $x_{ik'}$ are two covariates that are highly correlated, their contribution to the distances should be lower.
- Euclidean distance adjusted for covariance in the data
- Metric: $d(i,j) = \sqrt{(x_i - x_j)^T \hat{\Sigma}^{-1} (x_i - x_j)}$,

  where x_i and x_j are values of the matching variables for treated subject i and untreated subject j, and $\hat{\Sigma}$ is the sample covariance matrix of the observations

	Index	X1	X2
Treated	*i*	0	0
Control A	*A*	5	5
Control B	*B*	-4	0

$$\Sigma_{(X_1,X_2)} = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$$

Which control is closer to the i^{th} treated?

$$d(i,A) = \sqrt{(X_i - X_A)^T {\Sigma_X}^{-1} (X_i - X_A)}$$

$$= \sqrt{((0,0) - (5,5)) \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}^{-1} ((0,0) - (5,5)^T)}$$

$$= \sqrt{(-5,-5) \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} (-5,-5)^T} = 26$$

$$d(i,B) = \sqrt{(-4,0) \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} (-4,0)^T} = 84$$

Problem: As the number of dimensions in $\boldsymbol{x}$ increases, the average Euclidean or Mahalanobis distance between observations increases.

4. *Propensity Scores*

Propensity scores (PS): prob. of receiving the trtment given X_i

$$e(X_i) = P(A_i = 1|X_i)$$

– Reduce the dimension in X to one in PS

Balance Property: among those units with the same PS, X_i is identically distributed between the treated and untreated under conditional exchangeability and positivity: $X_i \perp A_i|e(X_i)$

- Absolute differences in the propensity score

$$d(i,j) = \left|e(X_i) - e\left(X_j\right)\right|$$

- Linear propensity score, $logit\left(e(X_i)\right) = X_i\beta$

$$d(i,j) = \left|logit\left(e(X_i)\right) - logit\left(e\left(X_j\right)\right)\right|$$

Accounting for nonlinearity in the substantive differences in PS

Prove: Balance Property $X_i \perp A_i | e(X_i)$

We know, to prove independence between A and B, all we need is to show Pr(A|B)=P(A):

$$\Pr(A_i = 1 | e(X_i), X_i) = E(A_i | e(X_i), X_i) = E(A_i | X_i) = P(A_i = 1 | X_i) = e(X_i)$$

We can also have:

$$\Pr\big(A_i = 1 \big| e(X_i)\big) = E\big(A_i \big| e(X_i)\big) = E[E(A_i | X_i) | e(X_i)] = E[e(X_i) | e(X_i)] = e(X_i) \text{ by law of iterative expectation}$$

Therefore, $\Pr(A_i = 1 | e(X_i), X_i) = \Pr(A_i = 1 | e(X_i))$, which implies the balancing property $A_i \perp X_i | e(X_i)$.

To have balanced covariate distribution, it is sufficient to just condition on $e(X_i)$, instead of high dimensional X_i!

We need to estimate $e(X_i)$ - under a model for a binary response!

Matching Methods

1. *Nearest Neighbor (NN) Matching*
   - Define $\boldsymbol{d}(.,.)$ be a metric between covariates x's
   - Let $I_t = \{1, 2, \dots, N_t\}$ be set of treated units
     Let $I_c = \{j(1), j(2), \dots, j(N_t)\}$ be set of matched controls
   - For each i, matching unit $j(i) = \underset{j \text{ s.t. } t_j \neq t_i}{argmin} d(x_j, x_i)$

     $j(i)$*:* the nearest counterfactual neighbor of the treated unit i in controls

- Implementation
  - Randomly ordering subjects in the treated group
  - Calculating the distance between the first treated and all untreated
  - The untreated subject, j, with the minimum distance $d(i,j)$ is the match for the treated subject i
  - Remove both subjects from the pool
  - Process is repeated until matches are found for all treated subjects.
- Problem: the order matters.
  For example (Treated PS: 0.5, 0.7; Control PS: 0.8, 0.15)
  - Match 0.5 first: 0.5 → 0.8, then 0.7 → 0.15 with $\sum d(i,j) = 0.3 + 0.55 = 0.85$
  - Match 0.7 first: 0.7→0.8, then 0.5→0.15 with $\sum d(i,j) = 0.45$

2. *Optimal matching*:
   Finds the matching solution that minimizes overall distance
   0.7→0.8 and 0.5→0.15

Assessing Balance

- Check covariant balance between treated and untreated after matching to determine whether the matching was successful

- Methods
  - Report standardized mean differences (whether the means of covariates are similar between the two groups)
  - Carry hypothesis testing such as 2-sample *t* test for continuous variable or chi-square test for discrete covariates (report p-value)

Standardized Difference

~ Difference in means between groups, divided by the (pooled) standard deviation

$$smd = \frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}},$$

where

$s^2_{treatment}$ sample variance in the treatment group for the covariate

$s^2_{control}$ sample variance in the control group for the covariate

- smd=1: there is a 1 standard deviation difference in x means

Key features of *smd*

- Scale invariant – age in years or in days would give same smd
- Doesn't depend on sample size
- Often, absolute value of *smd* is reported
- Calculate for each covariate that you match on
- Rules of thumb:
  - *smd* <0.1 indicate adequate balance
  - *smd* 0.1-0.2 are not too alarming
  - *smd* >0.2 indicate serious imbalance

How about two-sample t test? Depend on sample size

Table often appears in published paper

Summary statistics of main variables involved in the analyses

Example: right heart catheterization (RHC = 1 or 0)

	Unmatched			Matched		
	No RHC	RHC	SMD	No RHC	RHC	SMD
n	3551	2184		2082	2082	
age (mean (sd))	61.8 (17.3)	60.8 (15.6)	0.06	61.6 (16.7)	61.0 (15.8)	0.039
sex = Male (%)	53.9	58.5	0.09	56.9	56.9	0.001
resp = Yes (%)	41.7	28.9	0.27	30.6	30.4	0.005
card = Yes (%)	28.4	42.3	0.30	39.3	39.5	0.004
neuro = Yes (%)	16.2	5.4	0.35	5.3	5.7	0.015

Estimation of ACE after matching (Cont.):

Step 1: Match/Find each unit's counterfactual identical twin (as described above)

Step 2: Check on the outcome of treated and matched controls

y_i and $y_{j(i)}$ are outcomes of the treated and matched control

Step 3: Estimate both ACE and ICE

$$\widehat{ICE}(x_i) = \left(y_i - y_{j(i)}\right)$$

$$\widehat{ACE} = \frac{1}{n}\sum_{i=1:n} \widehat{ICE}(x_i)$$

Prove: Identification of ACE under Exact Matching

~ The distribution of X (matching variables) will be same for treated or matched controls: $\Pr(x|A=1) = \Pr(x|A=0, I_c)$

$E(Y^1 - Y^0)$

$= E(Y^1|A=1) - E(Y^0|A=1)$ by exchangeablity

$= E(Y|A=1) - E(Y^0|A=1)$ by consistency

$= E(Y|A=1) - \sum_x E(Y^0|A=1, X=x)\Pr(x|A=1)$ L.I.E.

$= E(Y|A=1) - \sum_x E(Y^0|A=0, X=x)\Pr(x|A=1)$ exchangeablity

$= E(Y|A=1) - \sum_x E(Y|A=0, X=x)\Pr(x|A=1)$ by consistency

$= E(Y|A=1) - \sum_x E(Y|A=0, X=x)\Pr(x|A=0, I_c)$ by exact match

The last row is identifiable!

R Codes – Matching

R packages: tableone and Matching

Functions:

CreateTableOne() in tableone

Match() in Matching

Matching Match()

```
greedymatch<-Match(Tr=treatment, M=1, X=mydata[xvars], replace=FALSE)
psmatch<-Match(Tr=mydata$treatment, M=1, X=logit(pscore),
                        replace=FALSE, caliper=0.2)
```

Tr	A vector indicating the observations, which are in the treatment. This can either be a logical vector or a real vector where 0 denotes control and 1 denotes treatment.
M	A scalar for the number of matches. The default is one-to-one matching.
X	A matrix containing the variables we wish to match on. This matrix may contain the actual observed covariates or the propensity score or a combination of both.
replace	A logical flag for whether matching should be done with replacement. Note that if FALSE, the order of matches generally matters. Matches will be found in the same order as the data are sorted. Thus, the match(es) for the first observation will be found first, the match(es) for the second observation will be found second, etc. Matching without replacement will generally increase bias.
caliper	A scalar or vector denoting the caliper(s), which should be used when matching. A caliper is the distance, which is acceptable for any match. Observations which are outside of the

caliper are dropped. If a scalar caliper is provided, this caliper is used for all covariates in X. If a vector of calipers is provided, a caliper value should be provided for each covariate in X. The caliper is interpreted to be in standardized units. For example, caliper=0.25 means that all matches >.25 standard deviations of each covariate in X are dropped. Note that dropping observations generally changes the quantity being estimated.

Outputs CreateTableOne()

table1<- CreateTableOne(vars=xvars, strata="treatment", data=mydata, test=FALSE)

vars Variables to be summarized given as a character vector. Factors are handled as categorical variables, whereas numeric variables are handled as continuous variables. If empty, all variables in the data frame specified in the data argument are used.

strata Stratifying (grouping) variable name(s) given as a character vector. If omitted, the overall results are returned.

data A data frame in which these variables exist. All variables (both vars and strata) must be in this data frame.

test If TRUE, as in the default and there are more than two groups, groupwise comparisons are performed.

print(table1,smd=TRUE)

The definitions of the standardized mean difference (SMD) are available in Flury *et al* 1986 for the univariate case and the multivariate case (essentially the square root of the Mahalanobis distance).

Right heart Catheterization (RHC) dataset

~This dataset was used in Connors et al. (1996): The effectiveness of RHC in the initial care of critically ill patients. J American Medical Association 276:889-897. The dataset pertains to day 1 of hospitalization, i.e., the "treatment" variable **swang1** is whether or not a patient received a RHC (also called the Swan-Ganz catheter) on the first day in which the patient qualified for the study. Connors et al. used binary logistic model to develop a propensity score that was then used for matching RHC patients with non-RHC patients.

```
####################
#RHC Example

#install packages
install.packages("tableone")
install.packages("Matching")

#load packages
library(tableone)
library(Matching)

#read in data
load("/datasets/rhc.sav")

#treatment variable is swang1: Primary disease category
#x variables that we will use
#cat1: primary disease category
```

```
#age
#sex
#meanbp1: mean blood pressure

#create a data set with just these variables, for simplicity
ARF<-as.numeric(rhc$cat1=='ARF')
CHF<-as.numeric(rhc$cat1=='CHF')
Cirr<-as.numeric(rhc$cat1=='Cirrhosis')
colcan<-as.numeric(rhc$cat1=='Colon Cancer')
Coma<-as.numeric(rhc$cat1=='Coma')
COPD<-as.numeric(rhc$cat1=='COPD')
lungcan<-as.numeric(rhc$cat1=='Lung Cancer')
MOSF<-as.numeric(rhc$cat1=='MOSF w/Malignancy')
sepsis<-as.numeric(rhc$cat1=='MOSF w/Sepsis')

female<-as.numeric(rhc$sex=='Female')
died<-as.numeric(rhc$death=='Yes')
age<-as.numeric(rhc$age)
treatment<-as.numeric(rhc$swang1=='RHC')
meanbp1<-rhc$meanbp1

#new dataset
mydata<-cbind(ARF,CHF,Cirr,colcan,Coma,lungcan,MOSF,sepsis, age,female,meanbp1,treatment,died)
mydata<-data.frame(mydata)
```

```
#covariates we will use (shorter list than you would use in practice)
xvars<-c("ARF","CHF","Cirr","colcan","Coma","lungcan","MOSF","sepsis", "age","female","meanbp1")

#look at a table 1
table1<- CreateTableOne(vars=xvars, strata="treatment", data=mydata, test=FALSE)
## include standardized mean difference (SMD)
print(table1,smd=TRUE)
```

	Stratified by treatment				
	0		1		SMD
n	3551		2184		
ARF (mean (sd))	0.45	0.5	0.42	0.49	0.059
CHF (mean (sd))	0.07	0.25	0.1	0.29	0.095
Cirr (mean (sd))	0.05	0.22	0.02	0.15	**0.145**
colcan (mean (sd))	0	0.04	0	0.02	0.038
Coma (mean (sd))	0.1	0.29	0.04	0.2	**0.207**
lungcan (mean (sd))	0.01	0.1	0	0.05	0.095
MOSF (mean (sd))	0.07	0.25	0.07	0.26	0.018
sepsis (mean (sd))	0.15	0.36	0.32	0.47	**0.415**
age (mean (sd))	61.76	17.29	60.75	15.63	0.061
female (mean (sd))	0.46	0.5	0.41	0.49	0.093
meanbp1 (mean (sd))	84.87	38.87	68.2	34.24	**0.455**

```
################################################
# do greedy matching on Mahalanobis distance
################################################
greedymatch<-Match(Tr=treatment, M=1, X=mydata[xvars], replace=FALSE)
matched<-mydata[unlist(greedymatch[c("index.treated","index.control")]), ]
#get table 1 for matched data with standardized mean differences
matchedtab1<-CreateTableOne(vars=xvars, strata ="treatment", data=matched, test = FALSE)
print(matchedtab1, smd = TRUE)
```

	Stratified by treatment				
	0		1		SMD
n	2184		2184		
ARF (mean (sd))	0.42	0.49	0.42	0.49	0.006
CHF (mean (sd))	0.1	0.29	0.1	0.29	<0.001

Cirr (mean (sd))	0.02	0.15	0.02	0.15	**<0.001**
colcan (mean (sd))	0	0.02	0	0.02	<0.001
Coma (mean (sd))	0.04	0.2	0.04	0.2	**<0.001**
lungcan (mean (sd))	0	0.05	0	0.05	<0.001
MOSF (mean (sd))	0.07	0.26	0.07	0.26	<0.001
sepsis (mean (sd))	0.24	0.43	0.32	0.47	**0.177**
age (mean (sd))	61.53	16.15	60.75	15.63	0.049
female (mean (sd))	0.44	0.5	0.41	0.49	0.042
meanbp1 (mean (sd))	73.12	34.28	68.2	34.24	**0.144**

```
#outcome analysis
y_trt<-matched$died[matched$treatment==1]
y_con<-matched$died[matched$treatment==0]

#pairwise difference
diffy<-y_trt-y_con
#paired t-test
t.test(diffy)
```

t = 3.9289, df = 2183, p-value = 8.799e-05

```
############################
# propensity score matching
############################
#fit a propensity score model - logistic regression
psmodel<-
glm(treatment~ARF+CHF+Cirr+colcan+Coma+lungcan+MOSF+sepsis+age+female+meanbp1,
family=binomial(),data=mydata)
```

```
#show coefficients etc
summary(psmodel)

#create propensity score
pscore<-psmodel$fitted.values

#do greedy matching on logit(PS) using Match with a caliper
logit <- function(p) {log(p)-log(1-p)}
psmatch<-Match(Tr=mydata$treatment, M=1, X=logit(pscore), replace=FALSE, caliper=.2)
matched<-mydata[unlist(psmatch[c("index.treated","index.control")]), ]
xvars<-c("ARF","CHF","Cirr","colcan","Coma","lungcan","MOSF","sepsis", "age","female","meanbp1")

#get standardized differences
matchedtab1<-CreateTableOne(vars=xvars, strata ="treatment", data=matched, test = FALSE)
print(matchedtab1, smd = TRUE)
```

	Stratified by treatment				
	0		1		SMD
n	1932		1932		
ARF (mean (sd))	0.47	0.5	0.47	0.5	0.002
CHF (mean (sd))	0.1	0.3	0.09	0.29	0.014
Cirr (mean (sd))	0.03	0.17	0.03	0.16	**0.019**
colcan (mean (sd))	0	0.04	0	0.02	0.032
Coma (mean (sd))	0.05	0.21	0.05	0.22	**0.010**
lungcan (mean (sd))	0	0.03	0	0.05	0.037
MOSF (mean (sd))	0.08	0.27	0.08	0.27	0.006

sepsis (mean (sd))	0.25	0.43	0.25	0.43	**0.004**
age (mean (sd))	61.05	17.82	60.91	15.52	0.009
female (mean (sd))	0.44	0.5	0.43	0.49	0.024
meanbp1 (mean (sd))	71.39	34.09	70.99	35.02	**0.012**

```
#outcome analysis
y_trt<-matched$died[matched$treatment==1]
y_con<-matched$died[matched$treatment==0]

#pairwise difference
diffy<-y_trt-y_con

#paired t-test
t.test(diffy)
t = 3.0396, df = 1931, p-value = 0.0046
```