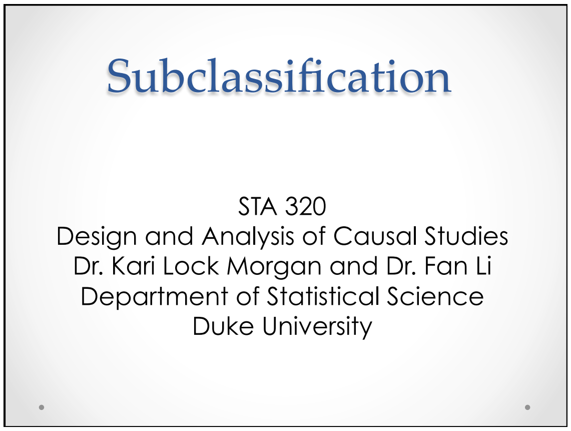
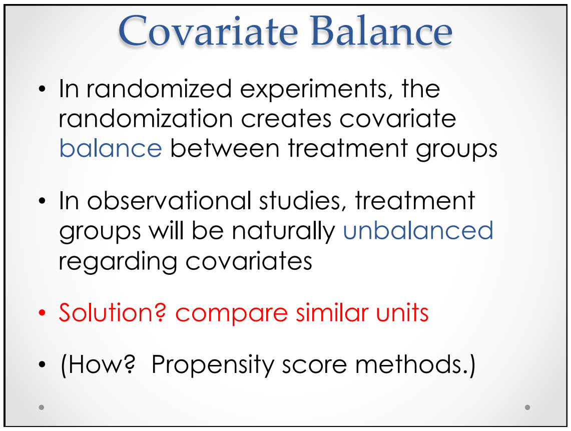


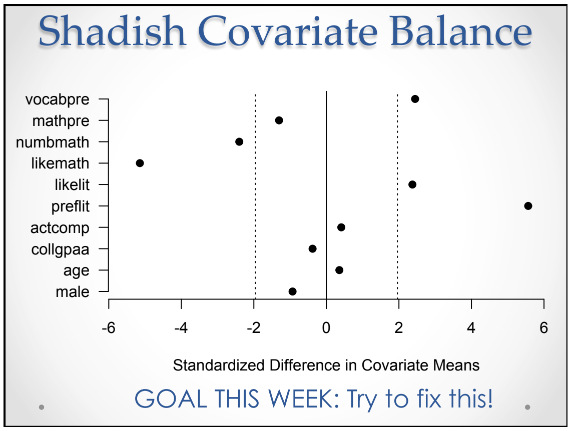
Subclassification

STA 320
Design and Analysis of Causal Studies
Dr. Kari Lock Morgan and Dr. Fan Li
Department of Statistical Science
Duke University



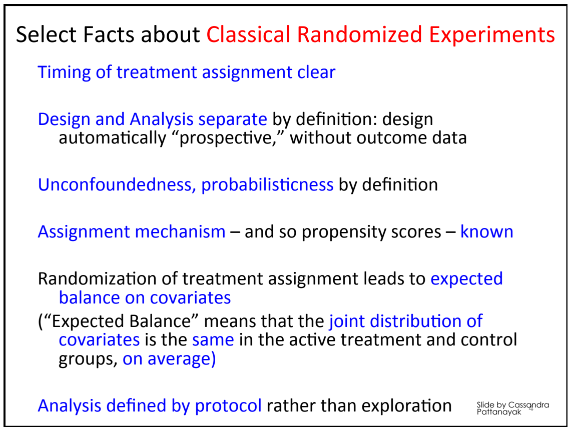
Covariate Balance

- In randomized experiments, the randomization creates covariate balance between treatment groups
- In observational studies, treatment groups will be naturally unbalanced regarding covariates
- Solution? compare similar units
- (How? Propensity score methods.)



Shadish Covariate Balance

GOAL THIS WEEK: Try to fix this!



Select Facts about Classical Randomized Experiments

Timing of treatment assignment clear

Design and Analysis separate by definition: design automatically "prospective," without outcome data

Unconfoundedness, probabilisticness by definition

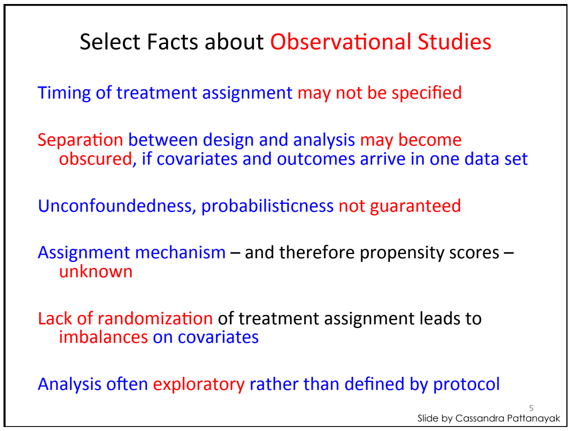
Assignment mechanism – and so propensity scores – known

Randomization of treatment assignment leads to expected balance on covariates

("Expected Balance" means that the joint distribution of covariates is the same in the active treatment and control groups, on average)

Analysis defined by protocol rather than exploration

Slide by Cassandra Pattanayak



Select Facts about Observational Studies

Timing of treatment assignment may not be specified

Separation between design and analysis may become obscured, if covariates and outcomes arrive in one data set

Unconfoundedness, probabilisticness not guaranteed

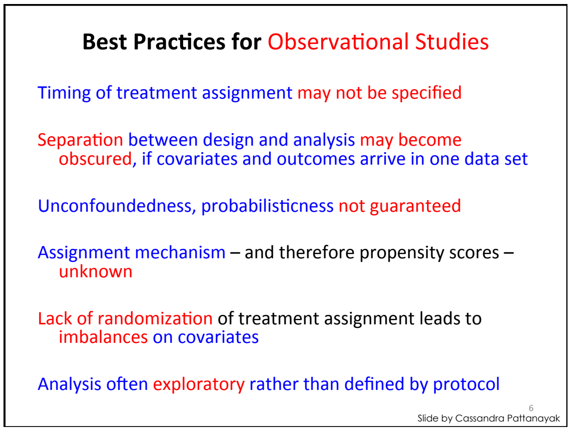
Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

5

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

Timing of treatment assignment may not be specified

Separation between design and analysis may become obscured, if covariates and outcomes arrive in one data set

Unconfoundedness, probabilisticness not guaranteed

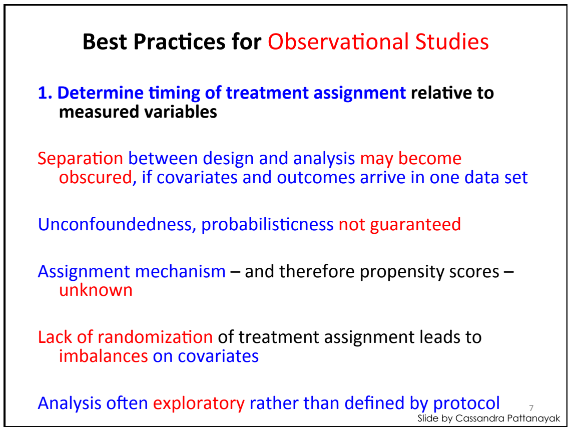
Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

6

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

Separation between design and analysis may become obscured, if covariates and outcomes arrive in one data set

Unconfoundedness, probabilisticness not guaranteed

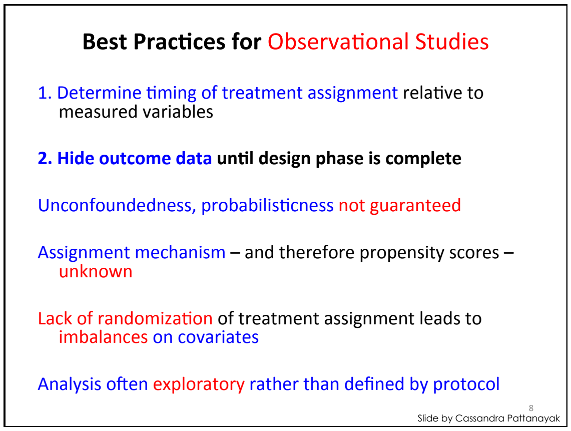
Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

7

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

Unconfoundedness, probabilisticness not guaranteed

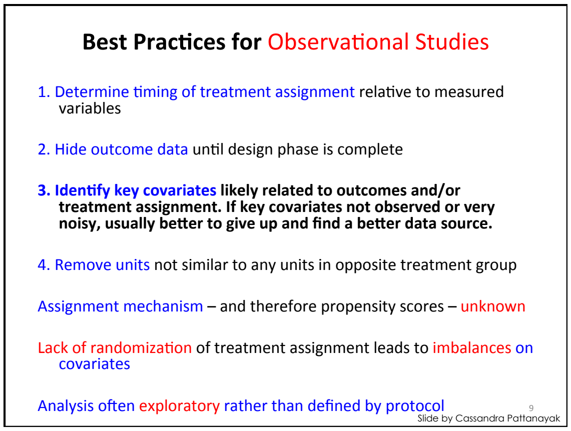
Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

8

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

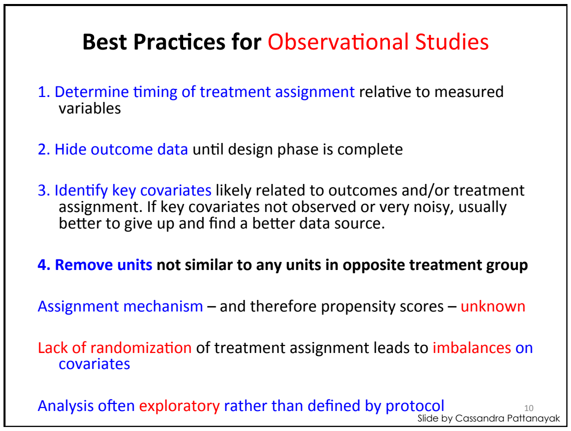
Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

9

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

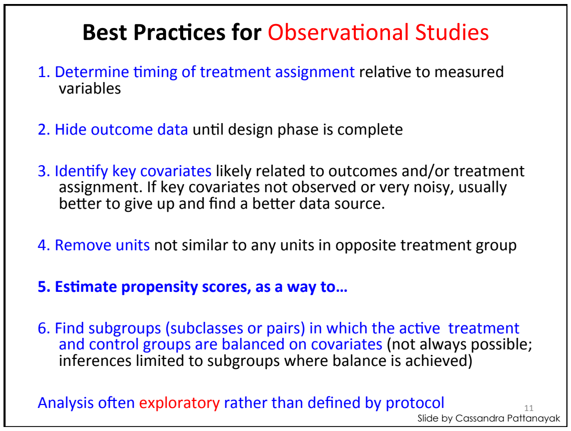
Assignment mechanism – and therefore propensity scores – unknown

Lack of randomization of treatment assignment leads to imbalances on covariates

Analysis often exploratory rather than defined by protocol

10

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

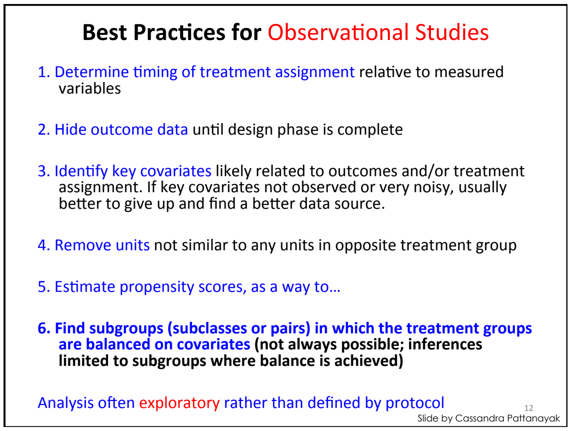
5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the active treatment and control groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

Analysis often exploratory rather than defined by protocol

11

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

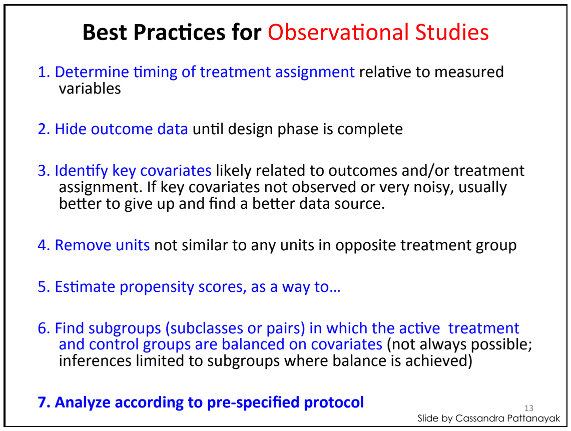
5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the treatment groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

Analysis often exploratory rather than defined by protocol

12

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

2. Hide outcome data until design phase is complete

3. Identify key covariates likely related to outcomes and/or treatment assignment. If key covariates not observed or very noisy, usually better to give up and find a better data source.

4. Remove units not similar to any units in opposite treatment group

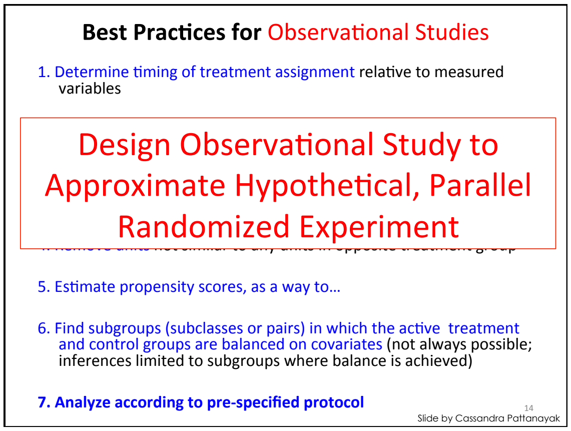
5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the active treatment and control groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

7. **Analyze according to pre-specified protocol**

13

Slide by Cassandra Pattanayak



Best Practices for Observational Studies

1. Determine timing of treatment assignment relative to measured variables

Design Observational Study to Approximate Hypothetical, Parallel Randomized Experiment

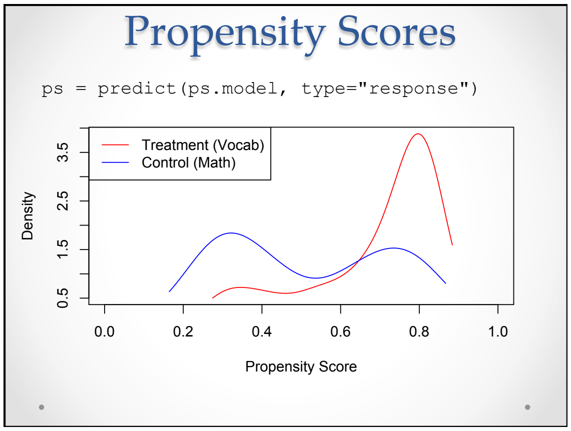
5. Estimate propensity scores, as a way to…

6. Find subgroups (subclasses or pairs) in which the active treatment and control groups are balanced on covariates (not always possible; inferences limited to subgroups where balance is achieved)

7. **Analyze according to pre-specified protocol**

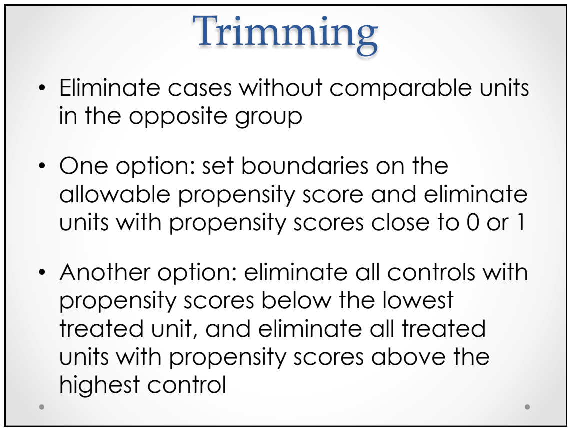
14

Slide by Cassandra Pattanayak



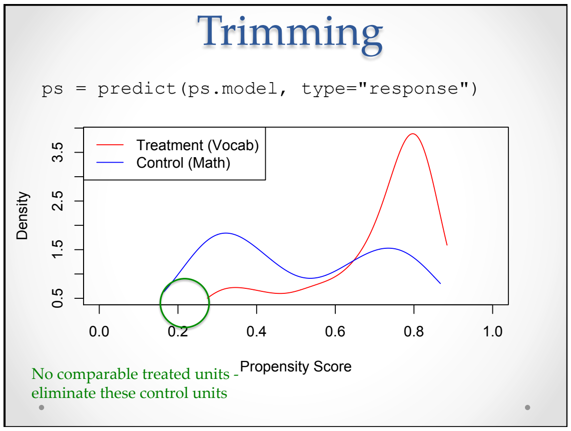
Propensity Scores

```
ps = predict(ps.model, type="response")
```



Trimming

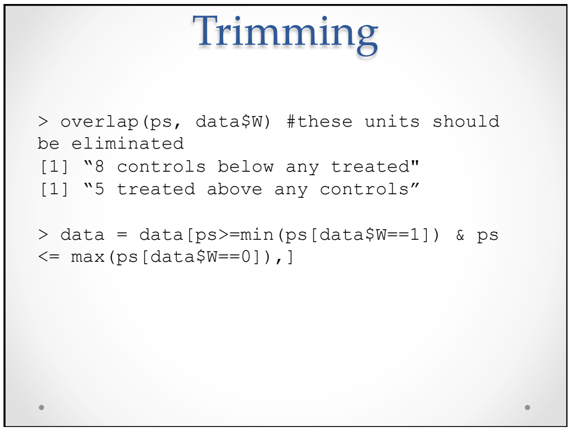
- Eliminate cases without comparable units in the opposite group
- One option: set boundaries on the allowable propensity score and eliminate units with propensity scores close to 0 or 1
- Another option: eliminate all controls with propensity scores below the lowest treated unit, and eliminate all treated units with propensity scores above the highest control



Trimming

```
ps = predict(ps.model, type="response")
```

No comparable treated units - eliminate these control units



Trimming

```
> overlap(ps, data$W) #these units should
be eliminated
[1] "8 controls below any treated"
[1] "5 treated above any controls"

> data = data[ps>=min(ps[data$W==1]) & ps
<= max(ps[data$W==0]),]
```

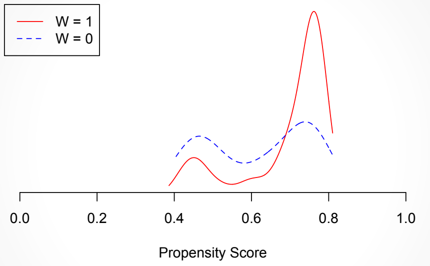
Estimating Propensity Scores

- In practice, estimating the propensity score is an iterative process:

1. Estimate propensity score
2. Eliminate units with no overlap (eliminate units with no comparable units in other groups)
3. Repeat until propensity scores overlapping everywhere for both groups

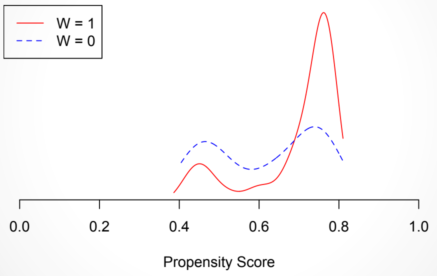
Go back and refit model

New Propensity Scores



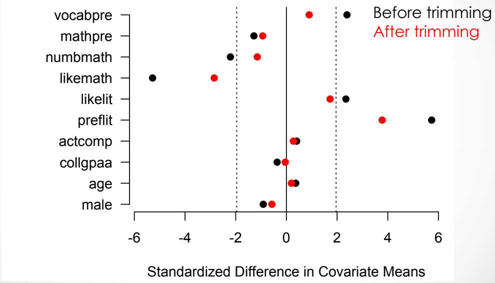
trim non-overlap…
refit propensity score model…

New Propensity Scores



After Trimming

- Original n = 210; after trimming n = 187



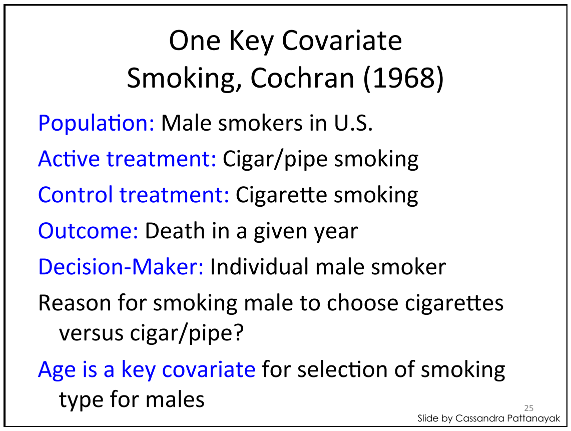
the closer to 0, the better! (0 = perfect balance)

Trimming

- Trimming can improve covariate balance, improving internal validity (better causal effects for remaining units)
- But hurts external validity (generalizability)
- Changes the estimand – estimate the causal effect for those units who are comparable
- How many units to trim is a tradeoff between decreasing sample size and better comparisons – Ch 16 gives optimal threshold

Subclasses

- If we have enough covariates (unconfounded), within subclasses of people with identical covariates, observational studies look like randomized experiments
- Idea: subclassify people based on similar covariate values, and estimate treatment effect within each subclass
- (similar to stratified experiments)



One Key Covariate Smoking, Cochran (1968)

Population: Male smokers in U.S.

Active treatment: Cigar/pipe smoking

Control treatment: Cigarette smoking

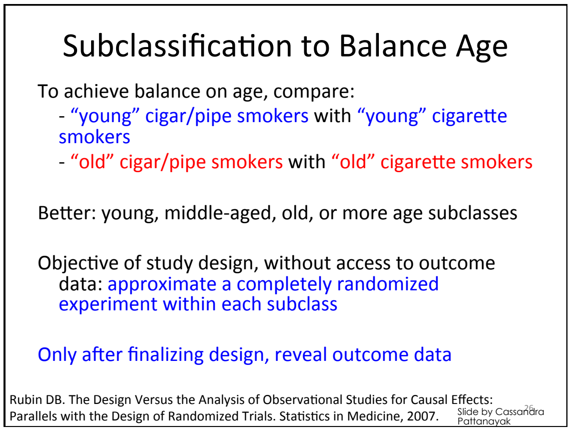
Outcome: Death in a given year

Decision-Maker: Individual male smoker

Reason for smoking male to choose cigarettes versus cigar/pipe?

Age is a key covariate for selection of smoking type for males

25 Slide by Cassandra Pattanayak



Subclassification to Balance Age

To achieve balance on age, compare:
- "young" cigar/pipe smokers with "young" cigarette smokers
- "old" cigar/pipe smokers with "old" cigarette smokers

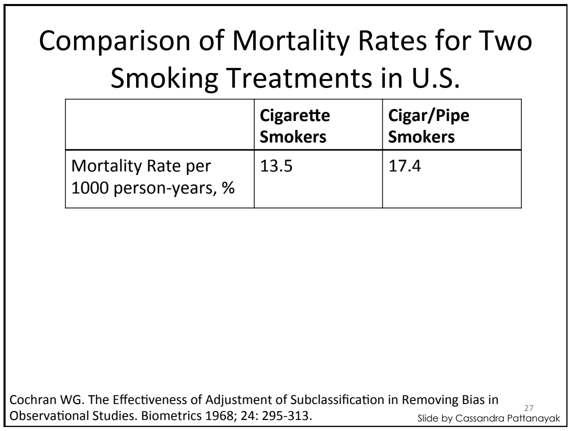
Better: young, middle-aged, old, or more age subclasses

Objective of study design, without access to outcome data: approximate a completely randomized experiment within each subclass

Only after finalizing design, reveal outcome data

Rubin DB. The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. Statistics in Medicine, 2007.

26 Slide by Cassandra Pattanayak

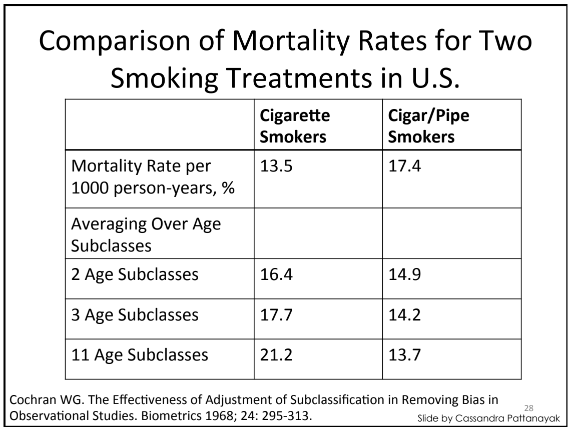


Comparison of Mortality Rates for Two Smoking Treatments in U.S.

	Cigarette Smokers	**Cigar/Pipe Smokers**
Mortality Rate per 1000 person-years, %	13.5	17.4

Cochran WG. The Effectiveness of Adjustment of Subclassification in Removing Bias in Observational Studies. Biometrics 1968; 24: 295-313.

27 Slide by Cassandra Pattanayak

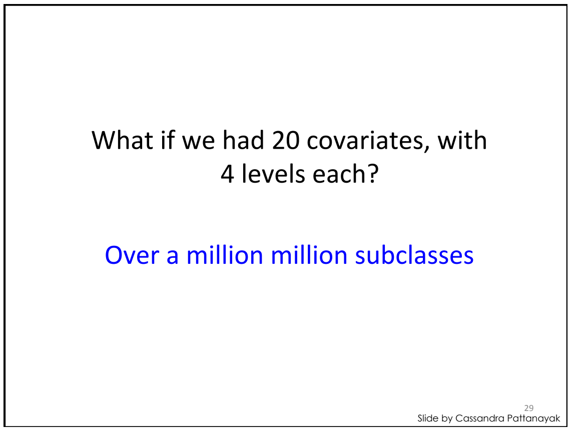


Comparison of Mortality Rates for Two Smoking Treatments in U.S.

	Cigarette Smokers	**Cigar/Pipe Smokers**
Mortality Rate per 1000 person-years, %	13.5	17.4
Averaging Over Age Subclasses		
2 Age Subclasses	16.4	14.9
3 Age Subclasses	17.7	14.2
11 Age Subclasses	21.2	13.7

Cochran WG. The Effectiveness of Adjustment of Subclassification in Removing Bias in Observational Studies. Biometrics 1968; 24: 295-313.

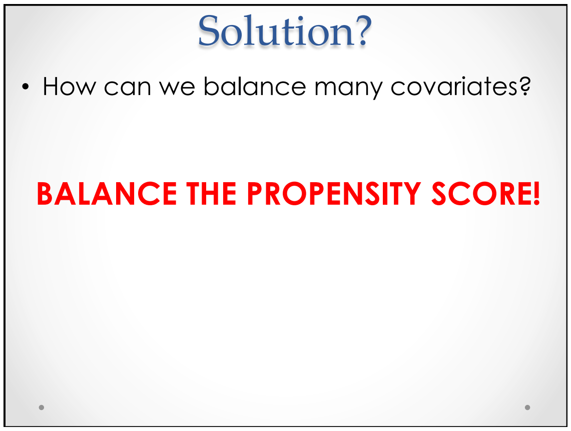
28 Slide by Cassandra Pattanayak



What if we had 20 covariates, with 4 levels each?

Over a million million subclasses

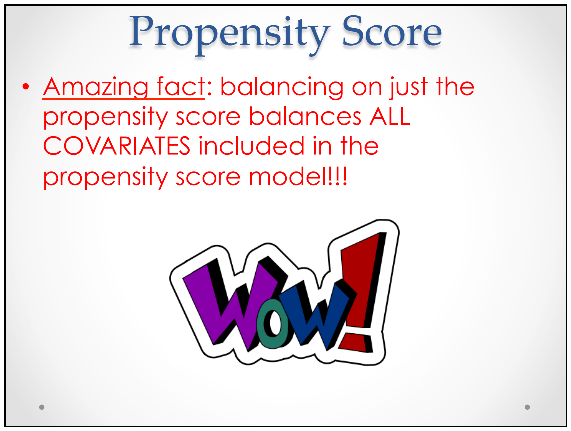
29 Slide by Cassandra Pattanayak



Solution?

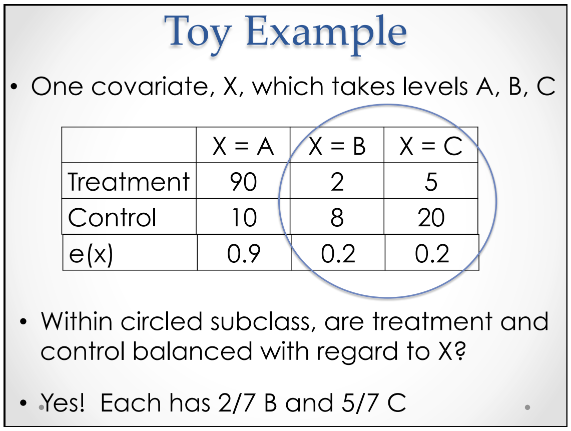
- How can we balance many covariates?

BALANCE THE PROPENSITY SCORE!



Propensity Score

- Amazing fact: balancing on just the propensity score balances ALL COVARIATES included in the propensity score model!!!

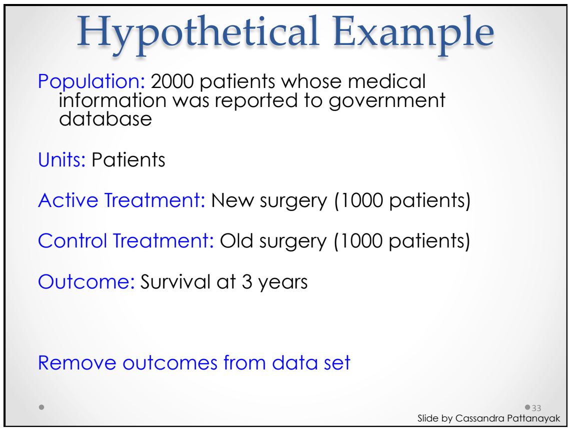


Toy Example

- One covariate, X, which takes levels A, B, C

	X = A	X = B	X = C
Treatment	90	2	5
Control	10	8	20
e(x)	0.9	0.2	0.2

- Within circled subclass, are treatment and control balanced with regard to X?
- Yes! Each has 2/7 B and 5/7 C



Hypothetical Example

Population: 2000 patients whose medical information was reported to government database

Units: Patients

Active Treatment: New surgery (1000 patients)

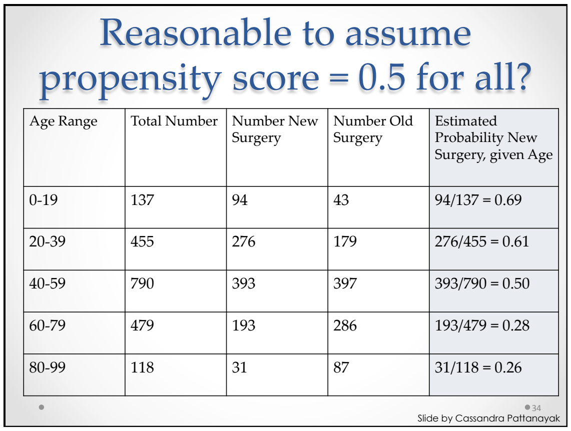
Control Treatment: Old surgery (1000 patients)

Outcome: Survival at 3 years

Remove outcomes from data set

33

Slide by Cassandra Pattanayak

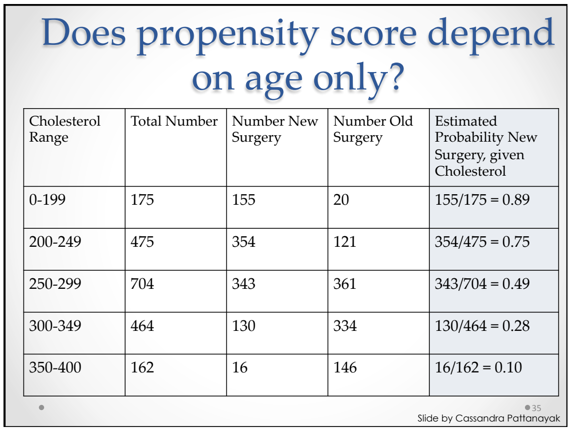


Reasonable to assume propensity score = 0.5 for all?

Age Range	Total Number	Number New Surgery	Number Old Surgery	Estimated Probability New Surgery, given Age
0-19	137	94	43	94/137 = 0.69
20-39	455	276	179	276/455 = 0.61
40-59	790	393	397	393/790 = 0.50
60-79	479	193	286	193/479 = 0.28
80-99	118	31	87	31/118 = 0.26

34

Slide by Cassandra Pattanayak

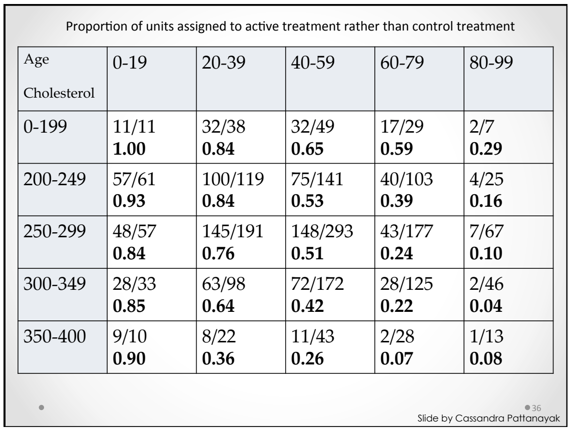


Does propensity score depend on age only?

Cholesterol Range	Total Number	Number New Surgery	Number Old Surgery	Estimated Probability New Surgery, given Cholesterol
0-199	175	155	20	155/175 = 0.89
200-249	475	354	121	354/475 = 0.75
250-299	704	343	361	343/704 = 0.49
300-349	464	130	334	130/464 = 0.28
350-400	162	16	146	16/162 = 0.10

35

Slide by Cassandra Pattanayak

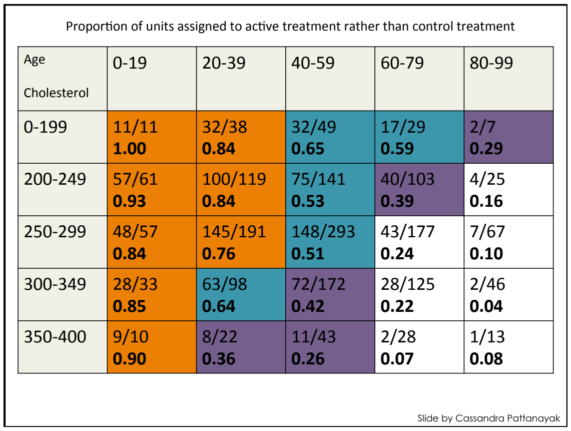


Proportion of units assigned to active treatment rather than control treatment

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	11/11 **1.00**	32/38 **0.84**	32/49 **0.65**	17/29 **0.59**	2/7 **0.29**
200-249	57/61 **0.93**	100/119 **0.84**	75/141 **0.53**	40/103 **0.39**	4/25 **0.16**
250-299	48/57 **0.84**	145/191 **0.76**	148/293 **0.51**	43/177 **0.24**	7/67 **0.10**
300-349	28/33 **0.85**	63/98 **0.64**	72/172 **0.42**	28/125 **0.22**	2/46 **0.04**
350-400	9/10 **0.90**	8/22 **0.36**	11/43 **0.26**	2/28 **0.07**	1/13 **0.08**

36

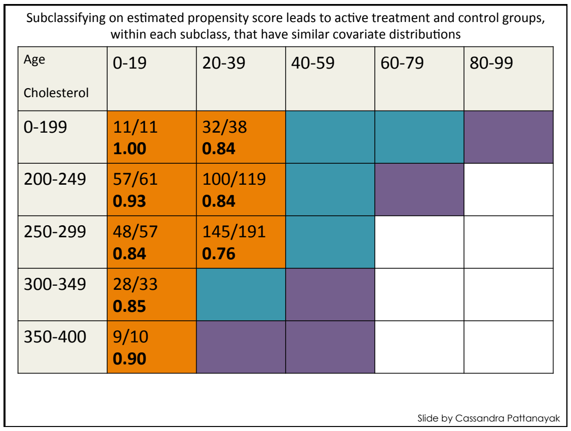
Slide by Cassandra Pattanayak



Proportion of units assigned to active treatment rather than control treatment

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	11/11 **1.00**	32/38 **0.84**	32/49 **0.65**	17/29 **0.59**	2/7 **0.29**
200-249	57/61 **0.93**	100/119 **0.84**	75/141 **0.53**	40/103 **0.39**	4/25 **0.16**
250-299	48/57 **0.84**	145/191 **0.76**	148/293 **0.51**	43/177 **0.24**	7/67 **0.10**
300-349	28/33 **0.85**	63/98 **0.64**	72/172 **0.42**	28/125 **0.22**	2/46 **0.04**
350-400	9/10 **0.90**	8/22 **0.36**	11/43 **0.26**	2/28 **0.07**	1/13 **0.08**

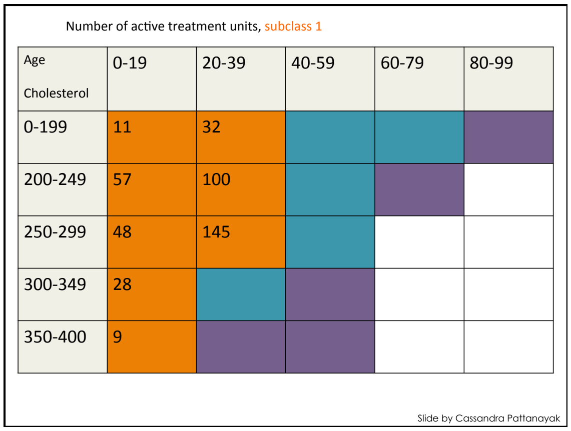
Slide by Cassandra Pattanayak



Subclassifying on estimated propensity score leads to active treatment and control groups, within each subclass, that have similar covariate distributions

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	11/11 **1.00**	32/38 **0.84**	🟦	🟦	🟪
200-249	57/61 **0.93**	100/119 **0.84**	🟦	🟪	
250-299	48/57 **0.84**	145/191 **0.76**	🟦		
300-349	28/33 **0.85**	🟦	🟪		
350-400	9/10 **0.90**	🟪	🟪		

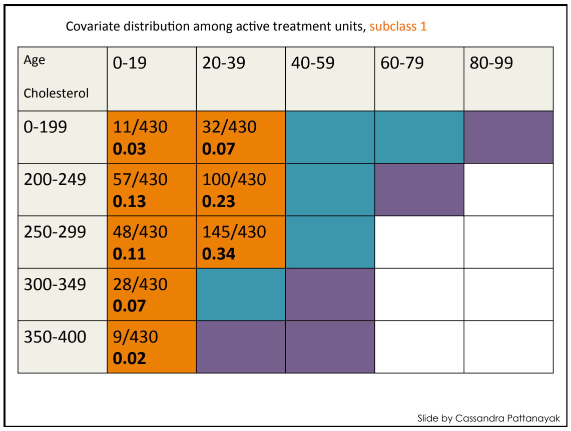
Slide by Cassandra Pattanayak



Number of active treatment units, subclass 1

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	11	32	🟦	🟦	🟪
200-249	57	100	🟦	🟪	
250-299	48	145	🟦		
300-349	28	🟦	🟪		
350-400	9	🟪	🟪		

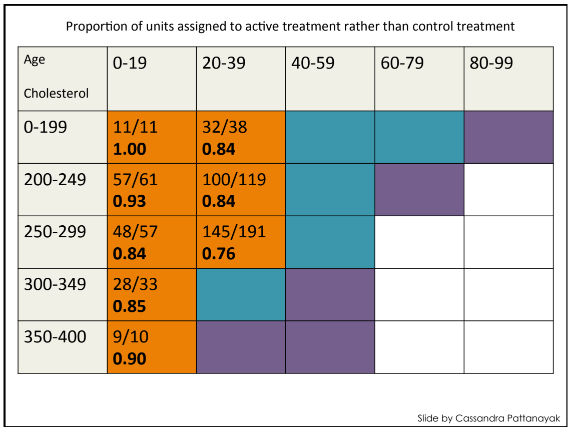
Slide by Cassandra Pattanayak



Covariate distribution among active treatment units, subclass 1

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	11/430 **0.03**	32/430 **0.07**	🟦	🟦	🟪
200-249	57/430 **0.13**	100/430 **0.23**	🟦	🟪	
250-299	48/430 **0.11**	145/430 **0.34**	🟦		
300-349	28/430 **0.07**	🟦	🟪		
350-400	9/430 **0.02**	🟪	🟪		

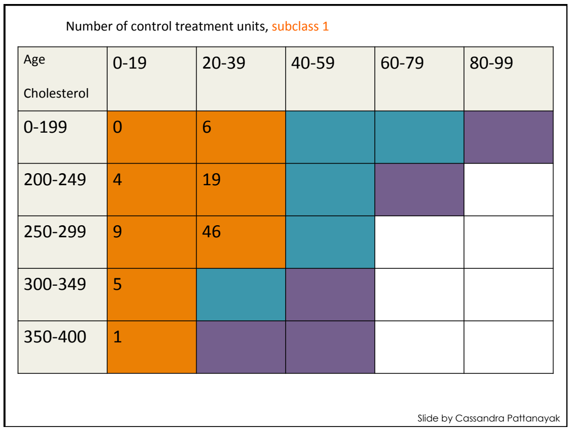
Slide by Cassandra Pattanayak



Proportion of units assigned to active treatment rather than control treatment

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	11/11 **1.00**	32/38 **0.84**	🟦	🟦	🟪
200-249	57/61 **0.93**	100/119 **0.84**	🟦	🟪	
250-299	48/57 **0.84**	145/191 **0.76**	🟦		
300-349	28/33 **0.85**	🟦	🟪		
350-400	9/10 **0.90**	🟪	🟪		

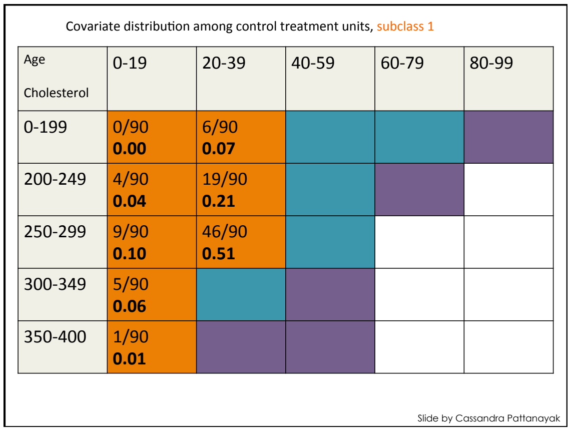
Slide by Cassandra Pattanayak



Number of control treatment units, subclass 1

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	0	6	🟦	🟦	🟪
200-249	4	19	🟦	🟪	
250-299	9	46	🟦		
300-349	5	🟦	🟪		
350-400	1	🟪	🟪		

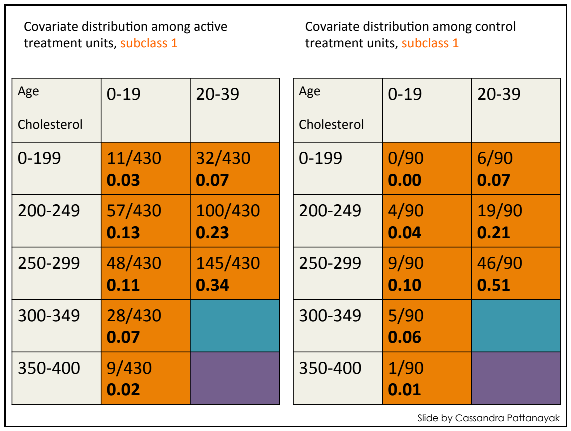
Slide by Cassandra Pattanayak



Covariate distribution among control treatment units, subclass 1

Age Cholesterol	0-19	20-39	40-59	60-79	80-99
0-199	0/90 **0.00**	6/90 **0.07**	🟦	🟦	🟪
200-249	4/90 **0.04**	19/90 **0.21**	🟦	🟪	
250-299	9/90 **0.10**	46/90 **0.51**	🟦		
300-349	5/90 **0.06**	🟦	🟪		
350-400	1/90 **0.01**	🟪	🟪		

Slide by Cassandra Pattanayak



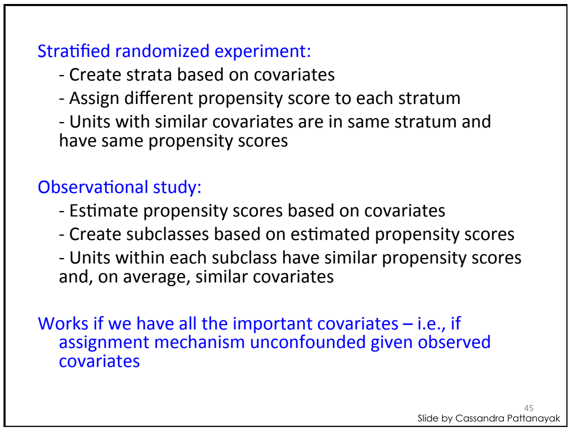
Covariate distribution among active treatment units, subclass 1

Age Cholesterol	0-19	20-39
0-199	11/430 **0.03**	32/430 **0.07**
200-249	57/430 **0.13**	100/430 **0.23**
250-299	48/430 **0.11**	145/430 **0.34**
300-349	28/430 **0.07**	🟦
350-400	9/430 **0.02**	🟪

Covariate distribution among control treatment units, subclass 1

Age Cholesterol	0-19	20-39
0-199	0/90 **0.00**	6/90 **0.07**
200-249	4/90 **0.04**	19/90 **0.21**
250-299	9/90 **0.10**	46/90 **0.51**
300-349	5/90 **0.06**	🟦
350-400	1/90 **0.01**	🟪

Slide by Cassandra Pattanayak



Stratified randomized experiment:

- Create strata based on covariates
- Assign different propensity score to each stratum
- Units with similar covariates are in same stratum and have same propensity scores

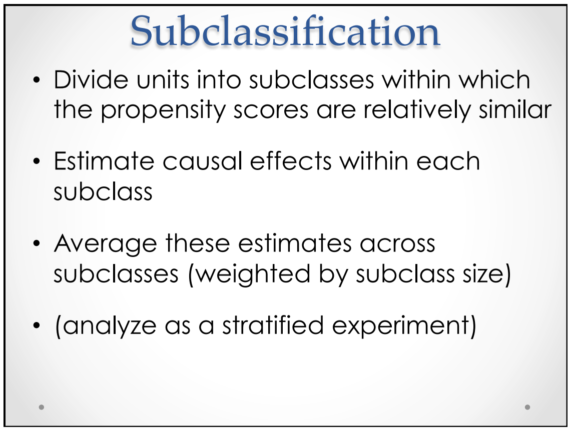
Observational study:

- Estimate propensity scores based on covariates
- Create subclasses based on estimated propensity scores
- Units within each subclass have similar propensity scores and, on average, similar covariates

Works if we have all the important covariates – i.e., if assignment mechanism unconfounded given observed covariates

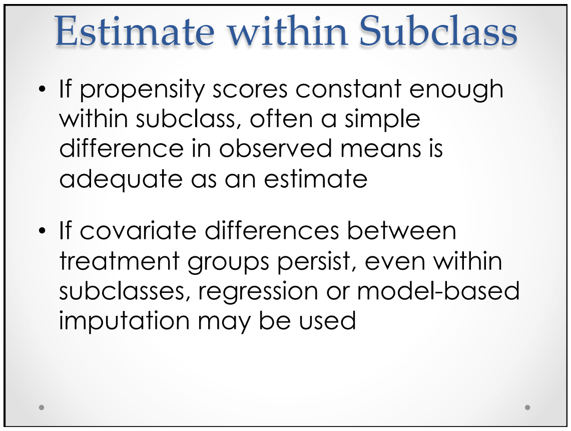
45

Slide by Cassandra Pattanayak



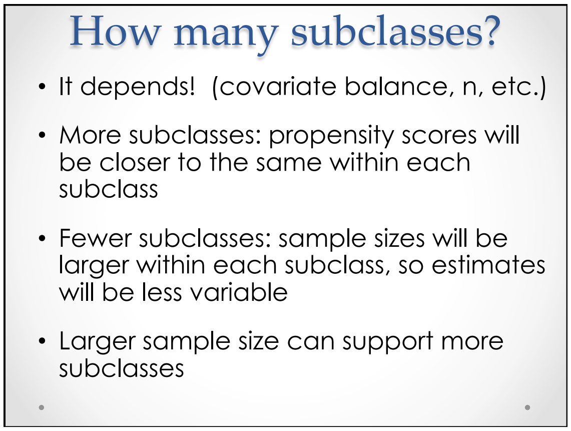
Subclassification

- Divide units into subclasses within which the propensity scores are relatively similar
- Estimate causal effects within each subclass
- Average these estimates across subclasses (weighted by subclass size)
- (analyze as a stratified experiment)



Estimate within Subclass

- If propensity scores constant enough within subclass, often a simple difference in observed means is adequate as an estimate
- If covariate differences between treatment groups persist, even within subclasses, regression or model-based imputation may be used



How many subclasses?

- It depends! (covariate balance, n, etc.)
- More subclasses: propensity scores will be closer to the same within each subclass
- Fewer subclasses: sample sizes will be larger within each subclass, so estimates will be less variable
- Larger sample size can support more subclasses

How many subclasses?

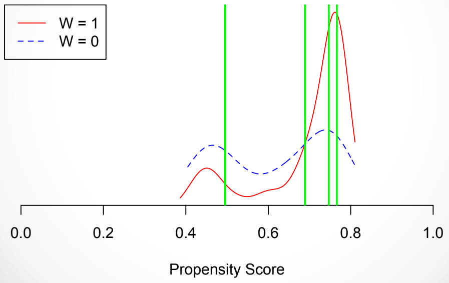
- Start with 5 equally sized subclasses (usually 5-10 are sufficient)
- Check…
  - Propensity score balance within subclasses
  - Number of treated and controls within subclasses
  - Overall covariate balance
- If balance needs improving and subclasses have enough treated and controls, try more subclasses

Subclass Breaks

- Starting with 5 equally sized subclasses
- Subclass breaks would be at the 20th, 40th, 60th, and 80th percentiles of the propensity score
- Subclasses do not have to be equally sized, that's just a convenient starting point

Shadish Data

```
> subclass.breaks = quantile(ps, c(.2,.4, .6,.8))
> subclass = subclasses(ps, subclass.breaks)
> plot.ps(ps, W, subclass.breaks)
```

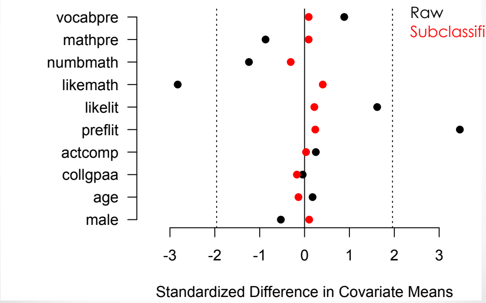


Shadish Data

```
> table(W, subclass)
   subclass
W    1  2  3  4  5
  0 19 18 12  7  7
  1 19 19 25 30 31
```

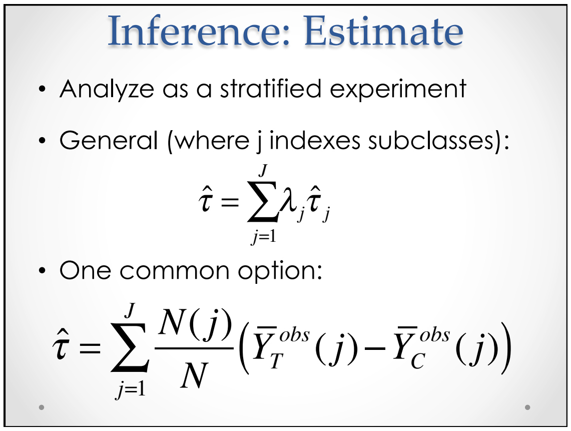
Shadish Data

```
> cov.balance(X, W)
> cov.balance(X, W, subclass)
```



Outcomes

- Once we are happy with the covariate balance, we can analyze the outcomes
- (Note: once you look at the outcomes, there is no turning back, so make sure you are happy with balance first!)



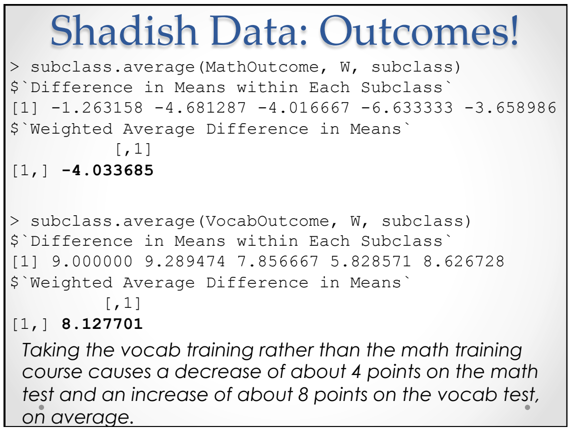
Inference: Estimate

- Analyze as a stratified experiment
- General (where j indexes subclasses):

$$\hat{\tau} = \sum_{j=1}^{J} \lambda_j \hat{\tau}_j$$

- One common option:

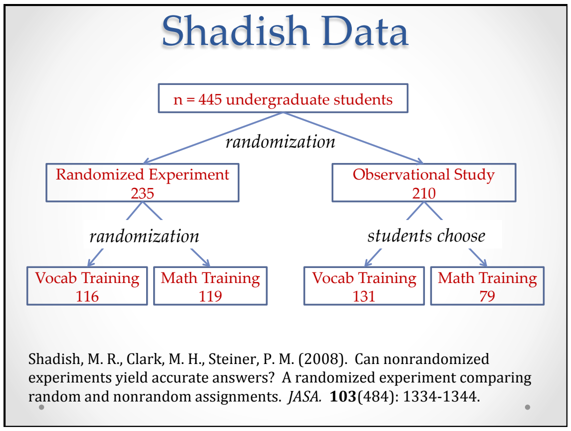
$$\hat{\tau} = \sum_{j=1}^{J} \frac{N(j)}{N} \left(\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j) \right)$$



Shadish Data: Outcomes!

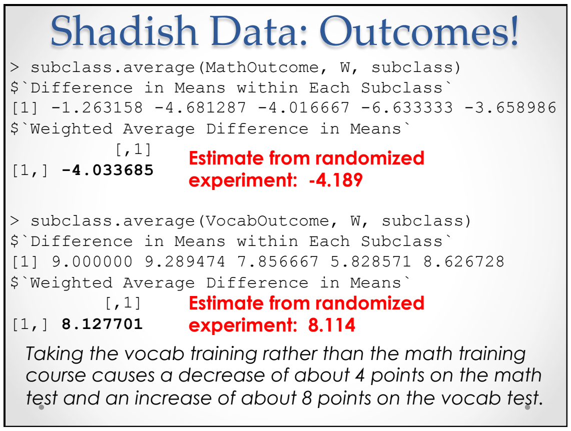
```
> subclass.average(MathOutcome, W, subclass)
$`Difference in Means within Each Subclass`
[1] -1.263158 -4.681287 -4.016667 -6.633333 -3.658986
$`Weighted Average Difference in Means`
          [,1]
[1,] -4.033685

> subclass.average(VocabOutcome, W, subclass)
$`Difference in Means within Each Subclass`
[1] 9.000000 9.289474 7.856667 5.828571 8.626728
$`Weighted Average Difference in Means`
         [,1]
[1,] 8.127701
```

Taking the vocab training rather than the math training course causes a decrease of about 4 points on the math test and an increase of about 8 points on the vocab test, on average.



Shadish Data

Shadish, M. R., Clark, M. H., Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *JASA.* **103**(484): 1334-1344.



Shadish Data: Outcomes!

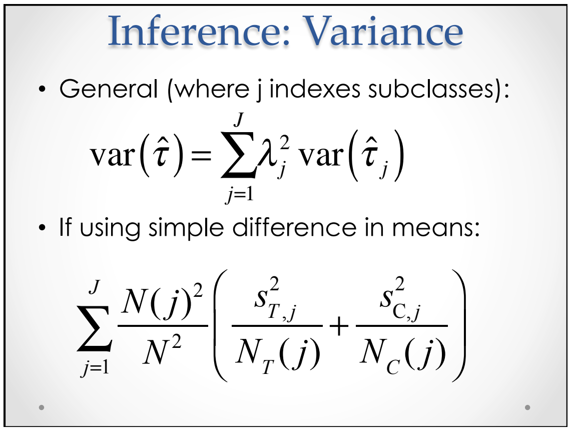
```
> subclass.average(MathOutcome, W, subclass)
$`Difference in Means within Each Subclass`
[1] -1.263158 -4.681287 -4.016667 -6.633333 -3.658986
$`Weighted Average Difference in Means`
          [,1]
[1,] -4.033685
```

Estimate from randomized experiment: -4.189

```
> subclass.average(VocabOutcome, W, subclass)
$`Difference in Means within Each Subclass`
[1] 9.000000 9.289474 7.856667 5.828571 8.626728
$`Weighted Average Difference in Means`
         [,1]
[1,] 8.127701
```

Estimate from randomized experiment: 8.114

Taking the vocab training rather than the math training course causes a decrease of about 4 points on the math test and an increase of about 8 points on the vocab test.



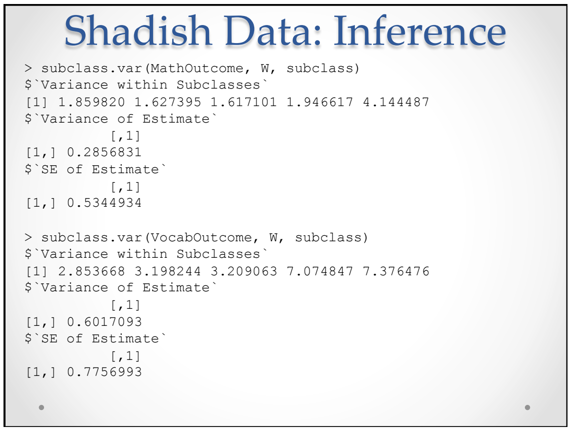
Inference: Variance

- General (where j indexes subclasses):

$$\operatorname{var}(\hat{\tau}) = \sum_{j=1}^{J} \lambda_j^2 \operatorname{var}(\hat{\tau}_j)$$

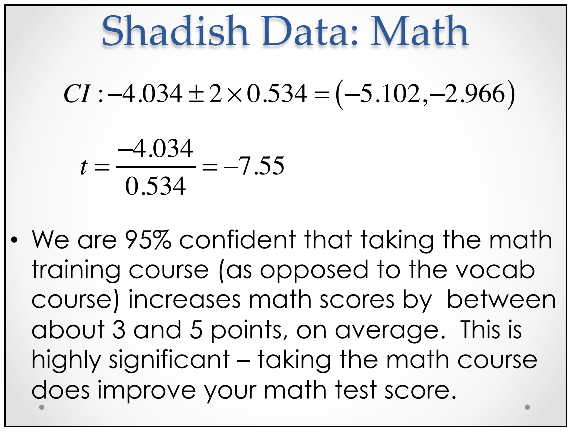
- If using simple difference in means:

$$\sum_{j=1}^{J} \frac{N(j)^2}{N^2} \left(\frac{s_{T,j}^2}{N_T(j)} + \frac{s_{C,j}^2}{N_C(j)} \right)$$



Shadish Data: Inference

```
> subclass.var(MathOutcome, W, subclass)
$`Variance within Subclasses`
[1] 1.859820 1.627395 1.617101 1.946617 4.144487
$`Variance of Estimate`
          [,1]
[1,] 0.2856831
$`SE of Estimate`
          [,1]
[1,] 0.5344934

> subclass.var(VocabOutcome, W, subclass)
$`Variance within Subclasses`
[1] 2.853668 3.198244 3.209063 7.074847 7.376476
$`Variance of Estimate`
          [,1]
[1,] 0.6017093
$`SE of Estimate`
          [,1]
[1,] 0.7756993
```

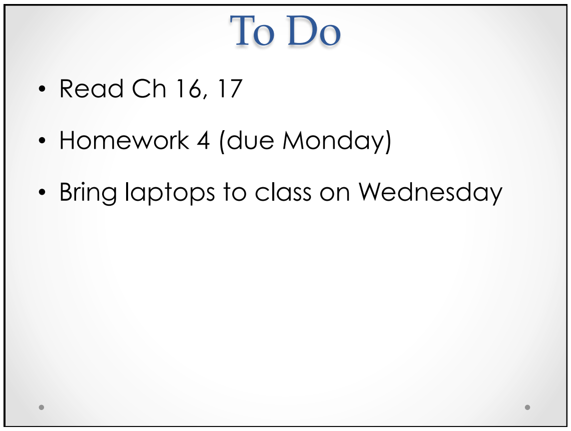


Shadish Data: Math

$$CI: -4.034 \pm 2 \times 0.534 = (-5.102, -2.966)$$

$$t = \frac{-4.034}{0.534} = -7.55$$

- We are 95% confident that taking the math training course (as opposed to the vocab course) increases math scores by between about 3 and 5 points, on average. This is highly significant – taking the math course does improve your math test score.



To Do

- Read Ch 16, 17
- Homework 4 (due Monday)
- Bring laptops to class on Wednesday