

Confounding and Effect Modification

Statistical Reasoning 2

Lecture 2

Lecture Topics

- Confounding
- Effect modification/statistical interaction

Section A

Confounding, an Introduction

Confounding (Lurking Variable)

- Consider results from the following (fictitious) study:
 - This study was done to investigate the association between smoking and a certain disease in male and female adults
 - 210 smokers and 240 non-smokers were recruited for the study

Results for	All Subjects		
	<i>Smokers</i>	<i>Non Smokers</i>	TOTALS
<i>Disease</i>	52	64	116
<i>No Disease</i>	158	176	334
TOTALS	210	240	450

$$\hat{RR} = \frac{\hat{p}_{smokers}}{\hat{p}_{non-smokers}} = \frac{52/210}{64/240} \approx 0.93$$

$$\hat{OR} = \frac{\hat{p}_{smokers} / (1 - \hat{p}_{smokers})}{\hat{p}_{non-smokers} / (1 - \hat{p}_{non-smokers})} = \frac{52 \times 176}{158 \times 64} \approx 0.91$$

What's Going On?

- Smoking is protective against disease?
- Most of the smokers are male and non-smokers are female

	All Subjects		
	<i>Smokers</i>	<i>Non Smokers</i>	TOTALS
<i>Male</i>	160	40	200
<i>Female</i>	50	200	250
TOTALS	210	240	450

5

What's Going On?

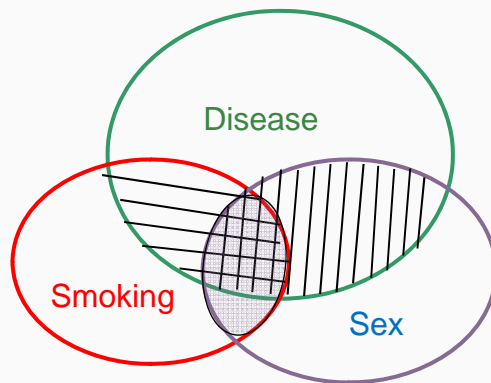
- Smoking is protective against disease?
- Further, most of the persons with disease are female

	All Subjects		
	<i>Disease</i>	<i>No Disease</i>	TOTALS
<i>Male</i>	33	167	200
<i>Female</i>	83	167	250
TOTALS	116	324	450

6

What's Going On?

- A picture?



7

What's Going On?

- The original outcome of interest is DISEASE
- The original exposure of interest is SMOKING
- In this sample, SEX is related to both the outcome and exposure
 - This relationship is possible impacting overall relationship between DISEASE and SMOKING
- How can we look at relationship between DISEASE and SMOKING removing any possible “interference” from SEX?
 - On approach - look at DISEASE and SMOKING relationship separately for males and females

8

Example

- Is smoking related to disease in males?

Results for	MALES		TOTALS
	<i>Smokers</i>	<i>Non Smokers</i>	
<i>Disease</i>	29	4	33
<i>No Disease</i>	131	36	167
TOTALS	160	40	200

$$\hat{RR}_{\text{males}} = \frac{\hat{p}_{\text{male smokers}}}{\hat{p}_{\text{male non-smokers}}} = \frac{29/160}{4/40} \approx 1.8$$

$$\hat{OR}_{\text{males}} = \frac{\hat{p}_{\text{male smokers}} / (1 - \hat{p}_{\text{male smokers}})}{p_{\text{male non-smokers}} / (1 - \hat{p}_{\text{male non-smokers}})} = \frac{29 \times 36}{4 \times 131} \approx 2$$

9

Example

- Is smoking related to disease in females?

Results for	FEMALES		TOTALS
	<i>Smokers</i>	<i>Non Smokers</i>	
<i>Disease</i>	23	60	83
<i>No Disease</i>	27	140	167
TOTALS	50	200	250

$$\hat{RR}_{\text{females}} = \frac{\hat{p}_{\text{female smokers}}}{\hat{p}_{\text{female non-smokers}}} = \frac{23/50}{60/200} \approx 1.5$$

$$\hat{OR}_{\text{females}} = \frac{\hat{p}_{\text{female smokers}} / (1 - \hat{p}_{\text{female smokers}})}{p_{\text{female non-smokers}} / (1 - \hat{p}_{\text{female non-smokers}})} = \frac{23 \times 140}{27 \times 60} \approx 2$$

10

Smoking, Disease and Sex

- A recap
 - The overall (sometimes called crude, unadjusted) relationship (RR) between smoking and disease was nearly 1 (risk difference nearly 0)

$$\hat{RR} = 0.93; \hat{p}_{smokers} - \hat{p}_{non-smokers} = -0.02$$

- The sex specific results showed similar positive associations between smoking and disease

$$\text{MALES: } \hat{RR} = 1.8; \hat{p}_{male\ smokers} - \hat{p}_{male\ non-smokers} \approx 0.08$$

$$\text{FEMALES: } \hat{RR} = 1.5; \hat{p}_{female\ smokers} - \hat{p}_{female\ non-smokers} \approx 0.16$$

(note, for the moment we are not considering statistical significance, just using estimates to illustrate point)

11

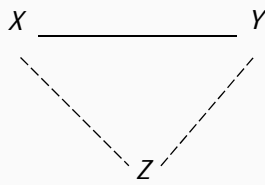
Simpson's Paradox

- The nature of an association can change (and even reverse direction) or disappear when data from several groups are combined to form a single group
- An association between an exposure **X** and a disease **Y** can be confounded by another lurking (hidden) variable **Z**

12

Confounding (Lurking Variable)

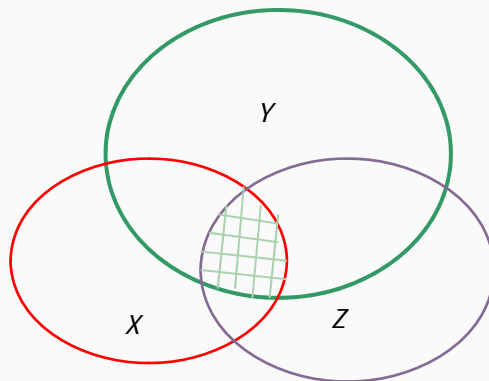
- A confounder Z distorts the true relation between X and Y
- This can happen if Z is related both to X and to Y



13

What's Going On?

- A picture



14

What is the Solution for Confounding?

- If you DON'T KNOW what the potential confounders are, there's not much you can do after the study is over
 - Randomization is the best protection
 - Randomization eliminates the potential links between the exposure of interest and potential confounders Z_1, Z_2, \dots, Z_3
- If you can't randomize but KNOW what the potential confounders are, or there are statistical methods to help control (adjust for confounders)
 - Potential confounders must be measured as part of study

15

How to Adjust for Confounding?

- Stratify
 - Look at tables separately
 - For example, male and females, clinic
 - Take weighted average of stratum specific estimates
- For example, in the disease/smoking situation
 - To get a sex adjusted relative risk for the smoking disease relationship we could weight the sex-specific relative risks by numbers of males and females

$$\hat{RR}_{sex\ adjusted} = \frac{n_{males} \times \hat{RR}_{males} + n_{females} \times \hat{RR}_{females}}{n_{males} + n_{females}}$$

$$\hat{RR}_{sex\ adjusted} = \frac{200 \times 1.8 + 250 \times 1.5}{200 + 250} \approx 1.6$$

16

How to Adjust for Confounding?

- There are better ways than this to take such a weighted average (weighting by standard error, for example), but this just illustrates the concept
- Confidence intervals can be computed for these adjusted measures of association
- One way to assess whether sex is a confounder: compare crude RR to sex adjusted RR: if “different” then sex is a confounder

17

How to Adjust for Confounding?

- Regression methods
 - Just around the corner!
 - More generalizable than weighted average approach, but idea is similar

18

Section B

Confounding, More Examples

Example 1: Arm Circumference and Height

- An observational study to estimate association between arm circumference and height in Nepali children
 - 94 randomly selected subjects, ages 3 months–6.5 years, had arm circumference, weight and height measured
 - This study is observational—it is not possible to randomize subjects to height groups!

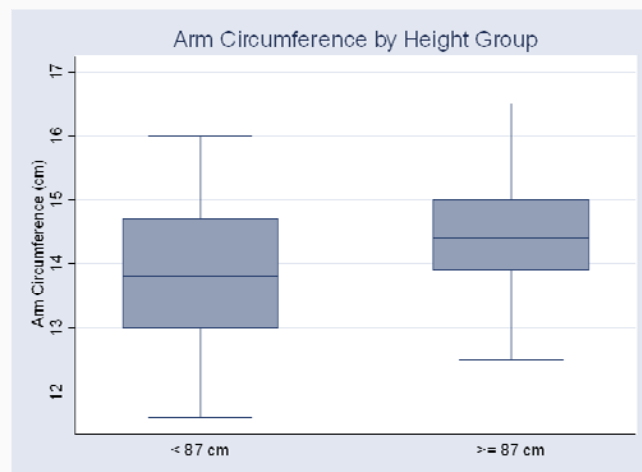
Example: Arm Circumference and Height

- The data
 - Arm circumference range: 11.6-16.5 cm
 - Height range: 57-109 cm
 - Weight range: 5-18 kg
- To perform analysis
 - Dichotomize height at median: i.e., subjects will be classified as “less than” or “greater than or equal to” median height of 87 cm
 - Dichotomize weight at median: i.e., subjects will be classified as “less than” or “greater than or equal to” median weight of 11.4 kg

21

Example: Arm Circumference and Height

- Boxplot arm circumference by height group



22

Example: Arm Circumference and Height

- Mean arm circumference (AC) by height group

Height Group	n	Mean AC	SD
< 87 cm	47	13.8	1.1
≥ 87 cm	47	14.5	0.9
Difference	-0.7 cm		

Shorter subjects have arm circumferences on average .7 cm lower than taller subjects (mean difference = -0.7, with 95% CI -1.1 cm to -0.3 cm)

23

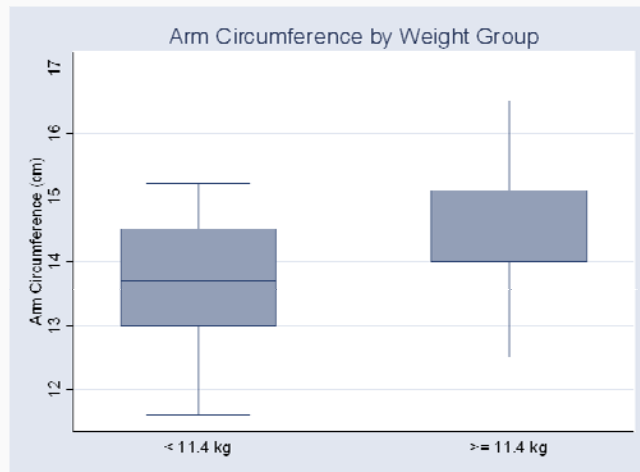
Example: Arm Circumference and Height

- However, it is very likely that arm circumference and height are both related to a child's weight
- Some of the relationship between arm circumference and height could be because of, or masked by these "behind the scenes" relationships to weight

24

Example: Arm Circumference and Height

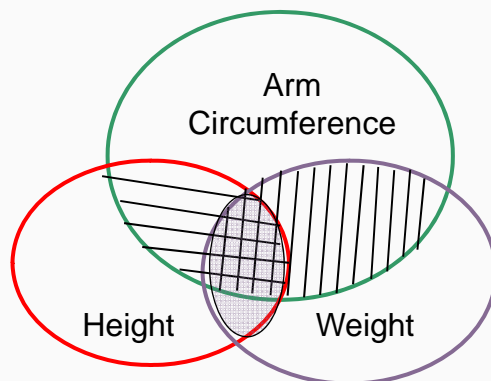
- What about weight?
 - Boxplot: arm circumference by weight group



25

Example: Arm Circumference and Height

- Possible diagram of this scenario



26

Example: Arm Circumference and Height

- Recall the original finding—children below the median height had arm circumferences of .7 cm lower on average than children (equal to or) above the median height
- To investigate whether this estimate is being fueled (or lessened) in part by weight differences in the height groups, and the arm circumference/weight relationship, let's stratify by weight group, and estimate the arm circumference/height association in each weight group

27

Example: Arm Circumference and Height

- Mean arm circumference (AC) by height group
 - Children below median weight

	<i>Lower Weight</i>	<i>Group</i>	
Height Group	n	Mean AC	SD
< 87 cm	41	13.65	1.1
≥ 87 cm	6	13.63	0.6

- Shorter subjects below the median weight have average arm circumferences on average .02 cm larger than taller subjects below the median weight (95% CI: - .64 cm (lower) to .68 cm (higher))

28

Example: Arm Circumference and Height

- Mean arm circumference (AC) by height group
 - Children above median weight

	<i>Higher Weight</i>	<i>Group</i>	
Height Group	n	Mean AC	SD
< 87 cm	6	14.65	0.92
≥ 87 cm	41	14.59	0.87

- Shorter subjects at or above the median weight have average arm circumferences on average .06 cm larger than taller subjects at or above the median weight (95% CI: - .90 cm (lower) to 1.0 cm (higher))

29

Example: Arm Circumference and Height

- A recap
 - Ignoring weight, children below the median height had arm circumferences of of .69 less on average than children at or above the median height and this difference was statistically significant
 - When stratified by weight children below the median height had arm circumferences marginally larger on average than children with or above the median height in both weight groups , but these estimates were very close to 0 and not statistically significant

30

Example: Arm Circumference and Height

- So, it appears as though the association between arm circumference and height “disappears” or at least gets much smaller after accounting for weight

- Associations: (mean difference in arm circumference, shorter subjects compared to taller)
 - Crude/unadjusted -0.7 cm (95% CI -1.1 to -0.3)
 - Adjusted?

one possibility: taking weighted average of weight specific AC/height associations, weighted by inverse of SE’s of weight specific associations

$$\frac{\frac{1}{SE(\bar{x}_{short\ low\ weight} - \bar{x}_{tall\ low\ weight})} \times (\bar{x}_{short\ low\ weight} - \bar{x}_{tall\ low\ weight}) + \frac{1}{SE(\bar{x}_{short\ high\ weight} - \bar{x}_{tall\ high\ weight})} \times (\bar{x}_{short\ high\ weight} - \bar{x}_{tall\ high\ weight})}{\frac{1}{SE(\bar{x}_{short\ low\ weight} - \bar{x}_{tall\ low\ weight})} + \frac{1}{SE(\bar{x}_{short\ high\ weight} - \bar{x}_{tall\ high\ weight})}}$$

31

Example: Arm Circumference and Height

- Associations: (mean difference in arm circumference, shorter subjects compared to taller)
 - Crude/unadjusted -0.7 cm (95% CI -1.1 to -0.3)
 - Adjusted?

one possibility: taking weighted average of weight specific associations, weighted by SE’s of weight specific associations

$$\frac{\frac{1}{0.46} \times .02 + \frac{1}{0.40} \times .06}{\frac{1}{0.46} + \frac{1}{0.40}} \approx 0.04$$

Can get 95% CI for this adjusted estimate: -0.30 cm to 0.38 cm

32

Example: Arm Circumference and Height

- One approach—take a weighted average of the average arm circumference differences between subjects below and above the median weight within weight groups, weighted by size of each group
- However, this is a pain, and if there are more potential confounders we could spend our life stratifying and computing such estimates
- Better approach—multiple regression methods (forthcoming!)

33

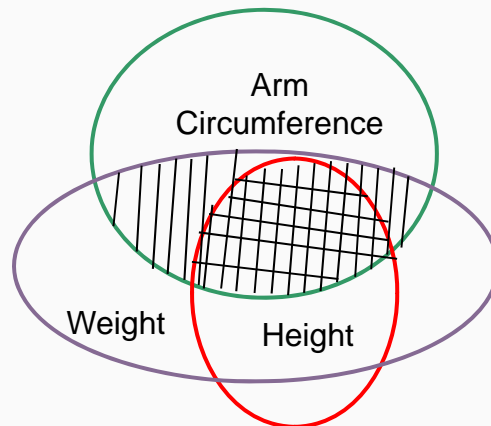
Example: Arm Circumference and Height

- Just FYI:
- A weighted overall average height adjusted difference in arm circumference between the two weight groups is .98 cm (children below median weight have smaller arm circumference on average), with 95% CI .40 cm to 1.55 cm
- Interesting:
 - When adjusted for weight, the arm circumference/height association disappears
 - When adjusted for height, the arm circumference/weight association is almost the same as the unadjusted arm circumference/weight association

34

Example: Arm Circumference and Height

- This is an interesting case, perhaps better illustrated by this picture:



35

Example: Arm Circumference and Height

- This is not always the case—many times when there is confounding between an outcome and two (or more) grouping variables, all of the adjusted outcome/group relationships will differ from the unadjusted associations

36

Example 2: South African Study

- A longitudinal study from South Africa: birth cohort, followed up five years after birth¹
- Participation by medical aid status at birth, all baseline participants

	All Subjects		
	Medical Aid	No Medical Aid	TOTAL
Follow-Up Participation	46	370	416
No Follow-Up Participation	195	979	1,164
TOTAL	241	1,349	1,590

$$\hat{RR}_{follow-up} = \frac{\hat{p}_{medical\ aid}}{\hat{p}_{no\ medical\ aid}} = \frac{46/241}{370/1,349} = \frac{0.19}{.27} \approx 0.70$$

37

Example 2: South African Study

- A longitudinal study from South Africa: birth cohort, followed up five years after birth
- Participation by medical aid status at birth, Black participants

	Black Subjects		
	Medical Aid	No Medical Aid	TOTAL
Follow-Up Participation	36	368	404
No Follow-Up Participation	91	957	1,048
TOTAL	127	1,325	1,452

$$\hat{RR}_{follow-up\ Black} = \frac{\hat{p}_{medical\ aid\ Black}}{\hat{p}_{no\ medical\ aid\ Black}} = \frac{36/127}{368/1,325} = \frac{0.28}{.28} \approx 1.0$$

38

Example 2: South African Study

- A longitudinal study from South Africa: birth cohort, followed up five years after birth
- Participation by medical aid status at birth, White participants

	White		
	Medical Aid	No Medical Aid	TOTAL
Follow-Up Participation	10	2	12
No Follow-Up Participation	104	22	126
TOTAL	114	24	138

$$\hat{RR}_{follow-up\ White} = \frac{\hat{p}_{medical\ aid\ White}}{\hat{p}_{no\ medical\ aid\ White}} = \frac{10/114}{2/24} = \frac{0.088}{.083} \approx 1.05$$

39

Example 2: South African Study

- Recap

40

Example 2: South African Study

- Whats going on?
- Race
 - Majority of sample Black subjects (91%)
- Race and follow-up participation
 - 26% of Black subjects completed follow-up as compared to 9% of White subjects
- Race and medical aid
 - 9% of Black subjects had medical aid compared to 83% of White subjects

41

Section C

Statistical Interaction / Effect Modification

Effect Modification / Interaction

- Let's look at the results from a fictitious data set comparing two treatments for a fatal disease as to the impact of each on reducing deaths: there are 600 “younger” patients and 600 “older” patients in this random sample of 1,200 patients
- Here is the data on all patients

	All Patients		
	Surgery	Drug	Total
Died	300	300	600
Survived	300	300	600
Total	600	600	

$$\hat{RR}_{death} = \frac{\hat{P}_{surgery}}{\hat{P}_{drug}} = \frac{300/600}{300/600} = 1; OR_{death} = 1$$

43

Effect Modification / Interaction

- Here is the data on only the 600 younger patients

	Younger	Patients	
	Surgery	Drug	Total
Died	100	200	300
Survived	200	100	300
Total	300	300	

$$\hat{RR}_{death\ younger} = \frac{\hat{P}_{surgery}}{\hat{P}_{drug}} = \frac{100/300}{200/300} = 0.5; OR_{death\ younger} = 0.25$$

44

Effect Modification / Interaction

- Here is the data on only the 600 older patients

	Older	Patients	
	Surgery	Drug	Total
Died	200	100	300
Survived	100	200	300
Total	300	300	

$$\hat{RR}_{\text{death older}} = \frac{\hat{P}_{\text{surgery}}}{\hat{P}_{\text{drug}}} = \frac{200/300}{100/300} = 2.0; O\hat{R}_{\text{death older}} = 4.0$$

45

Effect Modification / Interaction

- A recap of the overall and age specific results

46

Effect Modification / Interaction

- Age is an effect modifier:
 - Age modifies the association between death and treatment!
(Statistical interaction between age and treatment)
- The association between death and treatment depends on age
 - Surgery better for younger patients, drug therapy better for the older patients

47

Effect Modification / Interaction

- The association between death and one variable (TREATMENT) depends on the level of another variable (AGE)
- Here, it would not make sense to estimate one composite, overall measure of the association between death and treatment
- Best way to look at this data is to just look at the two tables (young and old) separately and estimate two separate death/treatment associations

48

Effect Modification / Interaction

- Here, it would not make sense to estimate one composite, overall measure of the association between death and treatment
- Best way to look at this data is to just look at the two tables (young and old) separately and estimate two separate death/treatment associations

49

Example: Tree Damage and Elevation

- Data on elevation and percentage of dead or badly damaged trees, from 64 Appalachian sites (reported by Committee on Monitoring and Assessment of Trends in Acid Deposition, 1986)
- Eight of the 64 sites are in Southern states
- Elevation information—whether the site was above or below 1,100 meters
- This is an observational study (why?)

50

Example: Tree Damage and Elevation

- Data for the first ten sites

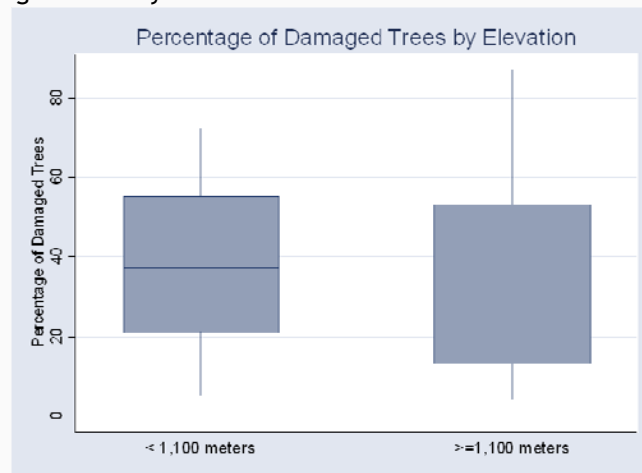
```
list damage elev_group region in 1/10
```

	damage	elev_group	region
1.	5	>=1,100 meters	South
2.	13	>=1,100 meters	South
3.	6	>-1,100 meters	South
4.	21	>-1,100 meters	South
5.	4	>=1,100 meters	South
6.	20	< 1,100 meters	South
7.	17	>-1,100 meters	South
8.	31	< 1,100 meters	South
9.	10	< 1,100 meters	North
10.	28	< 1,100 meters	North

51

Example: Tree Damage and Elevation

- Let's try to assess the relationship between the percentage of damaged trees and elevation—here is a boxplot of the percentage of damaged trees by elevation



52

Example: Tree Damage and Elevation

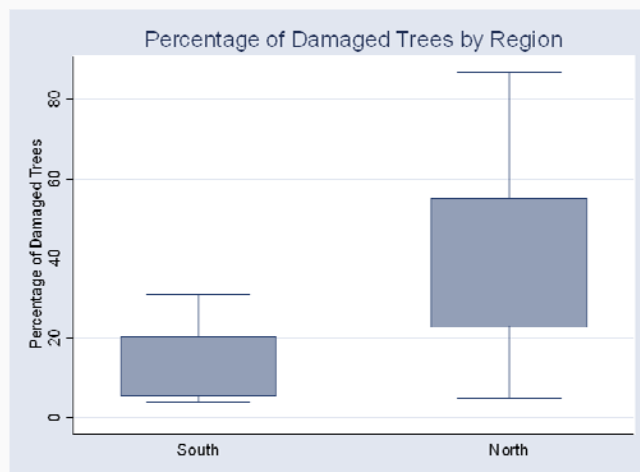
- Mean percentage of damaged trees by elevation group

Elevation	n	Mean Tree Damage (%)	SD
< 1,100 m	51	37.5	18.3
≥ 1,100 m	13	37.7	30.6

53

Example: Tree Damage and Elevation

- What about region though?
 - Boxplot percentage of damaged trees by region



54

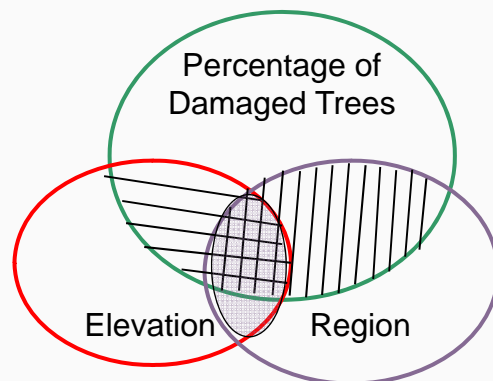
Example: Tree Damage and Elevation

- So sites in the South have less damage on average: i.e., not only is the percentage of damaged trees related to elevation, but it is also related to region
- If region is related to elevation, then it's possible that part of the relationship we saw (or didn't see) between damage and elevation is because of the region-damage-elevation relationship

55

Example: Tree Damage and Elevation

- Possible diagram of this scenario



56

Example: Tree Damage and Elevation

- Recall the original finding—sites with lower elevation had a marginally lower percentage of damaged trees on average: 0.2% less damaged than sites at higher elevation
- To adjust for regional differences in the elevation groups, and the damage/region relationship, let's stratify by region, and estimate the damage/elevation association in each region

57

Example: Tree Damage and Elevation

- Mean percentage of damage tree by elevation in Southern Sites

	<i>Southern</i>	<i>Sites Only</i>	
Elevation	n	Mean Tree Damage (%)	SD
< 1,100 m	2	25.5	7.7
≥ 1,100 m	6	11.0	7.1

- Southern sites at higher elevation have 14.5% less damaged trees on average than Southern sites at lower elevation (95% CI: 48.0 % lower–19.0 % higher)

58

Example: Tree Damage and Elevation

- Mean percentage of damage tree by elevation in Northern Sites

	<i>Northern</i>	<i>Sites Only</i>	
Elevation	n	Mean Tree Damage (%)	SD
< 1,100 m	49	38.0	18.5
≥ 1,100 m	7	61.0	22.6

- Northern sites at higher elevation have 23% greater damaged trees on average than Northern sites at lower elevation (95% CI: 2 % higher–44 % higher)

59

Example: Tree Damage and Elevation

- A recap
 - Ignoring region, sites with lower PCV sites with lower elevation had marginally lower percentage of damaged trees on average— 0.2 % less damaged than sites at higher elevation
- When stratified by region
 - Northern sites showed positive, statistically significant association between damage and elevation
 - Southern sites showed negative, non statistically-significant association between damage and elevation

60

Example: Tree Damage and Elevation

- So, it appears as though the association between tree damage and elevation is different, both in magnitude and direction depending on region
- We have a small dataset, so lack of statistical significance of negative damage/elevation associations in the South may be because of low power

61

Example: Tree Damage and Elevation

- One approach—take a weighted average of the average damage differences between sites at low and high elevations within each region, weighted by number of observations in each region
- However, does not necessarily make sense here—why combine estimates that differ in direction into one overall estimate?
- Better approach—to report two mean differences in damage between low and high elevation sites (one estimate for Northern sites, one estimate for Southern sites)
- Better approach—multiple regression methods (forthcoming!)

62