Confounding and Effect Modification

John McGready
Johns Hopkins University
Lecture Topics

Confounding

Effect modification/statistical interaction
Consider results from the following (fictitious) study:

- This study was done to investigate the association between smoking and a certain disease in males and female adults
- 210 smokers and 240 non-smokers were recruited for the study

We will first look at the results separately for males and females
Example

Is smoking related to disease in males?

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease</strong></td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td><strong>No Disease</strong></td>
<td>131</td>
<td>36</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>160</td>
<td>40</td>
</tr>
</tbody>
</table>

18% (29/160) of male smokers have disease
10% (4/40) of male non-smokers have disease

\[ \hat{RR} = 1.8, \hat{OR} = 2.0 \]
Is smoking related to disease in females?

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>23</td>
<td>60</td>
</tr>
<tr>
<td>No Disease</td>
<td>27</td>
<td>140</td>
</tr>
<tr>
<td>Totals</td>
<td>50</td>
<td>200</td>
</tr>
</tbody>
</table>

46% (23/50) of female smokers have disease
30% (60/200) of female non-smokers have disease

\[ \hat{RR} = 1.5, \hat{OR} = 2.0 \]
Example

Is smoking related to disease overall?

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease</strong></td>
<td>52</td>
<td>64</td>
</tr>
<tr>
<td><strong>No Disease</strong></td>
<td>158</td>
<td>176</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>210</td>
<td>240</td>
</tr>
</tbody>
</table>

25% (52/210) of smokers have disease
27% (64/240) of non-smokers have disease

\[ \hat{RR} = 0.93, \hat{OR} = 0.91 \]
There is a much higher prevalence of disease among females

<table>
<thead>
<tr>
<th>Disease and Sex</th>
<th>Disease</th>
<th>No Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>33</td>
<td>167</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>83</td>
<td>167</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>116</td>
<td>334</td>
</tr>
</tbody>
</table>
**What’s Going On?**

Most of the non-smokers are female

<table>
<thead>
<tr>
<th>Smoking and Sex</th>
<th>Smoker</th>
<th>Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>160</td>
<td>40</td>
</tr>
<tr>
<td>Female</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Totals</td>
<td>210</td>
<td>240</td>
</tr>
</tbody>
</table>

Continued
A picture

Disease

Smoking
What’s Going On?

A picture

Disease

Sex

Continued
A picture

What's Going On?

Smoking  Sex

Continued
A picture

Disease

Smoking

Sex

What’s Going On?
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 $Z$.
A confounder $Z$ distorts the true relation between $X$ and $Y$

This can happen if $Z$ is related both to $X$ and to $Y$
A picture
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*
If you can’t randomize but **KNOW** what the potential confounders are, or there are statistical methods to help control (adjust for confounders)
What is the Solution for Confounding?

Stratify

★ Look at tables separately
★ For example, male and females, clinic
★ Take weighted average of stratum specific estimates

Regression methods

★ Just around the corner!
Example: Arm Circumference and Height

An observational study to 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 weight 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
Example: Arm Circumference and Height

Boxplot arm circumference by height group

Arm Circumference by Height Group

< 87 cm  >= 87 cm
### Mean arm circumference (AC) by height group

<table>
<thead>
<tr>
<th>Height Group</th>
<th>n</th>
<th>Mean AC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 87 cm</td>
<td>47</td>
<td>13.8</td>
<td>1.1</td>
</tr>
<tr>
<td>≥ 87 cm</td>
<td>47</td>
<td>14.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Example: Arm Circumference and Height

Results from a t-test

```
. ttest armcirc, by( ht_group)

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 87 cm</td>
<td>47</td>
<td>13.78298</td>
<td>.1581378</td>
<td>1.084138</td>
<td>13.46466  14.10129</td>
</tr>
<tr>
<td>&gt;= 87 cm</td>
<td>47</td>
<td>14.47021</td>
<td>.1304271</td>
<td>.8941634</td>
<td>14.20768  14.73275</td>
</tr>
<tr>
<td>combined</td>
<td>94</td>
<td>14.1266</td>
<td>.1079877</td>
<td>1.04698</td>
<td>13.91215  14.34104</td>
</tr>
<tr>
<td>diff</td>
<td></td>
<td>-.687234</td>
<td>.2049849</td>
<td></td>
<td>-1.094352 -.2801163</td>
</tr>
</tbody>
</table>

Degrees of freedom: 92

Ho: mean(< 87 cm) - mean(>= 87 cm) = diff = 0

Ha: diff < 0  Ha: diff != 0  Ha: diff > 0
  t =  -3.3526  t =  -3.3526  t =  -3.3526
  P < t =  0.0006 P > |t| =  0.0012  P > t =  0.9994

Continued
Results from a t-test

- *Children below the median height had arm circumferences of .69 cm lower on average than children (equal to or) above the median height (95% CI: 0.28 cm–1.29 cm lower)*
What about weight?

- **Boxplot: arm circumference by weight group**
### Example: Arm Circumference and Height

Mean arm circumference (AC) by weight group

<table>
<thead>
<tr>
<th>Weight Group</th>
<th>n</th>
<th>Mean AC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 11.4 kg</td>
<td>47</td>
<td>13.6</td>
<td>1.0</td>
</tr>
<tr>
<td>≥ 11.4 kg</td>
<td>47</td>
<td>14.6</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Continued
### Example: Arm Circumference and Height

#### Results from a t-test

```
. ttest armcirc, by( wt_group)
```

Two-sample t test with equal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 11.4 k</td>
<td>47</td>
<td>13.65319</td>
<td>0.1468715</td>
<td>1.0069</td>
<td>13.35755 13.94883</td>
</tr>
<tr>
<td>&gt;= 11.4</td>
<td>47</td>
<td>14.6</td>
<td>0.1258459</td>
<td>0.8627559</td>
<td>14.34669 14.85331</td>
</tr>
<tr>
<td>combined</td>
<td>94</td>
<td>14.1266</td>
<td>0.1079877</td>
<td>1.04698</td>
<td>13.91215 14.34104</td>
</tr>
</tbody>
</table>

| diff | -0.9468085 | 0.1934126 | -1.330943 | -.5626745 |

Degrees of freedom: 92

Ho: mean(< 11.4 k) - mean(>= 11.4 ) = diff = 0

<table>
<thead>
<tr>
<th>Ha: diff &lt; 0</th>
<th>Ha: diff != 0</th>
<th>Ha: diff &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = -4.8953</td>
<td>t = -4.8953</td>
<td>t = -4.8953</td>
</tr>
<tr>
<td>P &lt; t = 0.0000</td>
<td>P &gt;</td>
<td>t</td>
</tr>
</tbody>
</table>

Continued
Results from a t-test

- **Children below the median weight had lower arm circumferences** .95 cm on average than children (equal to or) above the median weight (95% CI: .56 cm–1.33 cm lower)
Example: Arm Circumference and Height

So lower weight subjects have smaller arm circumferences on average: i.e., not only is arm circumference related to height, but it is also related to weight.

If height is related to weight, then it’s possible that part of the relationship we saw between arm circumference and height is because of the arm circumference-height-weight relationship—in other words, it is possible that the arm circumference/height relationship is confounded by weight.
Two-by-two table of height group versus weight group

<table>
<thead>
<tr>
<th>wt_group</th>
<th>ht_group</th>
<th>&lt; 87 cm</th>
<th>&gt;= 87 cm</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 11.4 kg</td>
<td>41</td>
<td>6</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>&gt;= 11.4 kg</td>
<td>6</td>
<td>41</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>47</td>
<td>94</td>
<td></td>
</tr>
</tbody>
</table>

Example: Arm Circumference and Height
87% of the heavier children are taller versus 13% of the younger subjects (difference = 74%, 95% CI 61%–88%)

So it appears that weight is also related to height
Example: Arm Circumference and Height

Possible diagram of this scenario
Example: Arm Circumference and Height

Recall the original finding—children below the median height had arm circumferences of .69 cm lower on average than children (equal to or) above the median height.

To investigate whether this estimate is being fueled 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.
Example: Arm Circumference and Height

Mean arm circumference (AC) by height group

- *Children below median weight*

<table>
<thead>
<tr>
<th>Height Group</th>
<th>n</th>
<th>Mean AC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 87 cm</td>
<td>41</td>
<td>13.65</td>
<td>1.1</td>
</tr>
<tr>
<td>≥ 87 cm</td>
<td>6</td>
<td>13.63</td>
<td>.62</td>
</tr>
</tbody>
</table>

- *Shorter subjects below the median weight have arm circumferences on average .02 cm larger than taller subjects below the median weight (95% CI: – .87 cm (lower) to .92 cm (higher))*
Example: Arm Circumference and Height

Mean arm circumference (AC) by height group

- *Children above median weight*

<table>
<thead>
<tr>
<th>Height Group</th>
<th>n</th>
<th>Mean AC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 87 cm</td>
<td>6</td>
<td>14.65</td>
<td>.92</td>
</tr>
<tr>
<td>≥ 87 cm</td>
<td>41</td>
<td>14.59</td>
<td>.87</td>
</tr>
</tbody>
</table>

- *Shorter subjects at or above the median weight have arm circumferences on average .06 cm larger than taller subjects at or above the median weight (95% CI: -.71 cm (lower) to .83 cm (higher))*
Example: Arm Circumference and Height

A recap

- Ignoring weight, children below the median height had arm circumferences of 0.69 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.

Continued
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
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.
Example: Arm Circumference and Height

Better approach—multiple regression methods (forthcoming!)
Example: Arm Circumference and Height

For completeness—let’s look at the relationship between arm circumference and weight, adjusted to height
Example: Arm Circumference and Height

Mean arm circumference (AC) by weight group

- **Children below median height**

<table>
<thead>
<tr>
<th>Weight Group</th>
<th>n</th>
<th>Mean AC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 11.4 kg</td>
<td>41</td>
<td>13.65</td>
<td>1.1</td>
</tr>
<tr>
<td>≥ 11.4 kg</td>
<td>6</td>
<td>14.65</td>
<td>.92</td>
</tr>
</tbody>
</table>

- **Lower weight children below the median height have arm circumferences on average one cm smaller than higher weight children below the median height (95% CI: .07 cm (lower) to 1.91 cm (lower))**
Example: Arm Circumference and Height

Mean arm circumference (AC) by weight group

★ *Children above median height*

<table>
<thead>
<tr>
<th>Weight Group</th>
<th>n</th>
<th>Mean AC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 11.4 kg</td>
<td>6</td>
<td>13.63</td>
<td>.62</td>
</tr>
<tr>
<td>≥ 11.4 kg</td>
<td>41</td>
<td>14.59</td>
<td>.89</td>
</tr>
</tbody>
</table>

★ *Lower weight children at or above the median height have arm circumferences on average .96 cm smaller than higher weight children at or above the median height (95% CI: .22 cm (lower to 1.7 cm (lower))*)
A recap

- Ignoring height, children below the median weight had arm circumferences of .95 cm less on average than children at or above the median height and this difference was statistically significant.
- When stratified by height children below the median weight had arm circumferences similarly smaller on average than children at or above the median height in both height groups, and these estimates were statistically significant.
Example: Arm Circumference and Height

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.
Example: Arm Circumference and Height

This is an interesting case, perhaps better illustrated by this picture:
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.
Section A

Practice Problems
1. Can you explain confounding to a classmate or other acquaintance? Can you think of an example to help with your explanation?
2. Consider a large, randomized study such as the Salk polio vaccine trial.
2. How does randomization help to ensure that the outcome/treatment group relationship observed is not confounded by any factors known or unknown? More specifically, what part of the necessary conditions for confounding does randomization eliminate?

Continued
3. Suppose a study were performed to assess the relationship between a diet choice (vegan, lacto-ovo vegetarian, neither) and cholesterol level. Subjects were not randomized to a dietary group. Before interpreting the relationship, can you name some potential factors for which it would be advisable to control?
Section A

Practice Problem Solutions
1. Can you explain confounding to a classmate or other acquaintance? Can you think of an example to help with your explanation?
2. Consider a large, randomized study such as the Salk polio vaccine trial.

School Children $\rightarrow$ Randomized

Vaccinated
N=200,745

Placebo
N=201,229

Continued
2. How does randomization help to ensure that the outcome/treatment group relationship observed is not confounded by any factors known or unknown? More specifically, what part of the necessary conditions for confounding does randomization eliminate?
Remember, a necessary condition for a variable to be a confounder is that it must be related to both the outcome and the predictor of interest.

In randomized studies, the “treatment” is the predictor of interest.

Randomization with a large number of potential subjects eliminates the association between the predictor of interest and any other variable besides (potentially) the outcome.
In other words, randomization allows the only difference between the “treatment” groups to be the treatment of interest.
In the “confounding conditions” diagram with outcome $Y$, predictor of interest $X$, and potential confounder $Z$: 

```
X ---- Z ---- Y
    |        |        |
      |        |        |
        |        |        |
      |        |        |
        |        |        |
          Y
```
Randomization eliminates the necessary link between $X$ and $Z$
3. Suppose a study were performed to assess the relationship between a diet choice (vegan, lacto-ovo vegetarian, neither) and cholesterol level. Subjects were not randomized to a dietary group. Before interpreting the relationship, can you name some potential factors for which it would be advisable to control?
Remember, for a factor to be a confounder, it must be associated with both cholesterol level and with being vegetarian.
Some candidates include the following:

- **Gender**—females are more likely to be vegetarian and have lower cholesterol
- **Smoking**—vegetarians are less likely to smoke and smoking is associated with higher cholesterol
- **Income**—vegetarians are wealthier and increased wealth associated with lower cholesterol
Section B

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

- The data presented is stratified by two age groups, “young” and “old”
The data:

<table>
<thead>
<tr>
<th></th>
<th>YOUNG</th>
<th></th>
<th>OLD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surgery</td>
<td>Drug</td>
<td>Surgery</td>
<td>Drug</td>
</tr>
<tr>
<td>Died</td>
<td>100</td>
<td>200</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Survived</td>
<td>200</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Totals</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>
Effect Modification / Interaction

Surgery is better for this group!

<table>
<thead>
<tr>
<th></th>
<th>Young</th>
<th>Old</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surgery</td>
<td>Drug</td>
</tr>
<tr>
<td>Died</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Survived</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Totals</td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>

33% (100/300) of those who had surgery died, as compared to 67% (200/300) of those taking the drug!

(RR = .50, OR = .25)
Taking the drug is better for this group!

<table>
<thead>
<tr>
<th></th>
<th>YOUNG</th>
<th></th>
<th>OLD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surgery</td>
<td>Drug</td>
<td>Surgery</td>
<td>Drug</td>
</tr>
<tr>
<td>Died</td>
<td>100</td>
<td>200</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Survived</td>
<td>200</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Totals</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>

67% (200/300) of those who had surgery died, as compared to 37% (100/300) of those who took the drug!

(RR = 2, OR = 4)
What happens when we combine tables?

<table>
<thead>
<tr>
<th></th>
<th>Combined Young and Old</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surgery</td>
</tr>
<tr>
<td>Died</td>
<td>300</td>
</tr>
<tr>
<td>Survived</td>
<td>300</td>
</tr>
<tr>
<td>Totals</td>
<td>600</td>
</tr>
</tbody>
</table>

Surgery and drug groups have identical proportions dying! (50% in each group!)
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**
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
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?)
Example: Tree Damage and Elevation

Data for the first ten sites

```
list damage elev_group region in 1/10

+----------------------------------+
<table>
<thead>
<tr>
<th>damage       elev_group   region</th>
</tr>
</thead>
</table>
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 |
+----------------------------------+

Continued
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.
### Example: Tree Damage and Elevation

Mean percentage of damaged trees by elevation group

<table>
<thead>
<tr>
<th>Elevation Group</th>
<th>n</th>
<th>Tree Damage(%)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1,100 m</td>
<td>51</td>
<td>37.5</td>
<td>18.3</td>
</tr>
<tr>
<td>≥ 1,100 m</td>
<td>13</td>
<td>37.7</td>
<td>30.6</td>
</tr>
</tbody>
</table>
Example: Tree Damage and Elevation

Results from a t-test

```
. ttest damage, by( elev_group) unequal

Two-sample t test with unequal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1,100</td>
<td>51</td>
<td>37.4902</td>
<td>2.56559</td>
<td>18.32198</td>
<td>[32.33706, 42.64334]</td>
</tr>
<tr>
<td>&gt;=1,100</td>
<td>13</td>
<td>37.69231</td>
<td>8.496141</td>
<td>30.63327</td>
<td>[19.18081, 56.20381]</td>
</tr>
</tbody>
</table>

combined | 64  | 37.53125| 2.637392  | 21.09914  | [32.26084, 42.80166] |

diff | -.2021116| 8.875059 | [-19.20474, 18.80051] |

Satterthwaite's degrees of freedom: 14.2598

Ho: mean(< 1,100 ) - mean(>=1,100 ) = diff = 0

Ha: diff < 0  Ha: diff != 0  Ha: diff > 0
  t = -0.0228    t = -0.0228    t = -0.0228
  P < t = 0.4911  P > |t| = 0.9821  P > t = 0.5089
```

Continued
Example: Tree Damage and Elevation

Results from a t-test

- Sites with lower elevation had marginally lower percentage of damaged trees on average: 0.2% less damaged than sites at higher elevation (95% CI: 19.2 % lower—18.8 % higher)
Example: Tree Damage and Elevation

What about region though?

★ Boxplot percentage of damaged trees by region
### Example: Tree Damage and Elevation

Mean percentage of damaged trees by region

<table>
<thead>
<tr>
<th>Region</th>
<th>$n$</th>
<th>Mean Tree Damage(%)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>8</td>
<td>14.6</td>
<td>9.5</td>
</tr>
<tr>
<td>North</td>
<td>56</td>
<td>40.8</td>
<td>20.3</td>
</tr>
</tbody>
</table>
### Example: Tree Damage and Elevation

Results from a t-test

```plaintext
. ttest damage, by( region) unequal
Two-sample t test with unequal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>8</td>
<td>14.625</td>
<td>3.343103</td>
<td>9.455724</td>
<td>6.719817 - 22.53018</td>
</tr>
<tr>
<td>North</td>
<td>56</td>
<td>40.80357</td>
<td>2.71018</td>
<td>20.28113</td>
<td>35.37225 - 46.23489</td>
</tr>
<tr>
<td>combined</td>
<td>64</td>
<td>37.53125</td>
<td>2.637392</td>
<td>21.09914</td>
<td>32.26084 - 42.80166</td>
</tr>
</tbody>
</table>

diff | -26.17857 | 4.303652 | -35.2123 - 17.14484

Satterthwaite’s degrees of freedom: 18.2225

Ho: mean(South) - mean(North) = diff = 0

Ha: diff < 0    Ha: diff != 0    Ha: diff > 0
  t = -6.0829    t = -6.0829    t = -6.0829
  P < t = 0.0000 P > |t| = 0.0000 P > t = 1.0000

.
```

Continued
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.
### Example: Tree Damage and Elevation

Two-by-two table of elevation group by region

```
. tab region elev_group

<table>
<thead>
<tr>
<th>region</th>
<th>&lt; 1,100 m</th>
<th>&gt;=1,100 m</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>North</td>
<td>49</td>
<td>7</td>
<td>56</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>13</td>
<td>64</td>
</tr>
</tbody>
</table>
```

*Continued*
Example: Tree Damage and Elevation

75% of the sites in the south are at higher elevation as compared to 13% of the northern sites (difference = 62%, 95% CI 31%–94%)

So it appears that region is also related to elevation
Example: Tree Damage and Elevation

Possible diagram of this scenario
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.
Example: Tree Damage and Elevation

Mean percentage of damage tree by elevation stratified by region

- **South**

<table>
<thead>
<tr>
<th>Elevation Group</th>
<th>n</th>
<th>Mean Tree Damage(%)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1,100 m</td>
<td>2</td>
<td>25.5</td>
<td>7.7</td>
</tr>
<tr>
<td>≥ 1,100 m</td>
<td>6</td>
<td>11.0</td>
<td>7.1</td>
</tr>
</tbody>
</table>

- **Southern sites at lower elevation have 14.5% more damaged trees on average than Southern sites at high elevation (95% CI: 19.7 % lower—48.7 % higher)**
Example: Tree Damage and Elevation

Mean percentage of damage tree by elevation stratified by region

- **North**

<table>
<thead>
<tr>
<th>Elevation Group</th>
<th>n</th>
<th>Tree Damage(%)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1,100 m</td>
<td>49</td>
<td>38.0</td>
<td>18.5</td>
</tr>
<tr>
<td>≥ 1,100 m</td>
<td>7</td>
<td>60.6</td>
<td>22.6</td>
</tr>
</tbody>
</table>

- *Northern sites at lower elevation have 22.6% less damaged trees on average than Northern sites at high elevation (95% CI: 1.5% lower—43.6 % lower)*
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, significant association between damage and elevation**
- **Southern sites showed negative, non-significant association between 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.
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!)
Section B

Practice Problems
1. Explain the difference between confounding and statistical interaction.

2. Can you envision a design situation where there is no confounding but there is interaction?
1. Explain the difference between confounding and statistical interaction.

★ A lot of people have trouble with these two concepts, as they are difficult. Confounding occurs when the relationship between two factors (exposure and outcome) is obscured by a third, hidden factor.
1. Explain the difference between confounding and statistical interaction.

- **Statistical interaction** occurs when the relationship between two factors (exposure and outcome) is different across different levels of a third factor. It is possible to uncover both confounding and statistical interaction at the same time.
2. Can you envision a design situation where there is no confounding, but there is interaction?

- *We will discuss this further in LiveTalk—but feel free to add your thoughts on this to a BBS topic.*