# Confounding Variables

Ryan Miller

# Introduction

- *Statistical inference* provides us the tools to identify whether an observed relationship might be explained by *chance*
  - However, a small $p$-value *does not* imply the relationship is *causal*

# Introduction

- *Statistical inference* provides us the tools to identify whether an observed relationship might be explained by *chance*
  - However, a small *p*-value *does not* imply the relationship is *causal*
- When arguing for a *cause-effect relationship*, we must be able to rule out *all* other possible explanations (in addition to chance)

# Study Design

*Study design* refers to the way data are collected. There are two major categories of study design:

1) **Observational designs** - the data are simply observed/recorded without any active involvement by the researcher
2) **Experimental designs** - the researcher actively influences the explanatory variable of interest

A very important type of experimental design is the *randomized experiment*

# Observational Data

- Gender bias is a long-standing issue in higher education
- In 1975, statisticians at UC-Berkley analyzed graduate admissions data for UC-Berkley
  - Overall, 1195 of 2691 (44.5%) male applicants were accepted, while only 557 of 1835 (30.4%)
  - Statistically speaking, is this a compelling difference?

# Observational Data

- Gender bias is a long-standing issue in higher education
- In 1975, statisticians at UC-Berkley analyzed graduate admissions data for UC-Berkley
  - Overall, 1195 of 2691 (44.5%) male applicants were accepted, while only 557 of 1835 (30.4%)
  - Statistically speaking, is this a compelling difference?

```r
## Hypothesis Test in R
prop.test(x = c(1198,557), n = c(2691, 1835))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(1198, 557) out of c(2691, 1835)
## X-squared = 91.61, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1129887 0.1703022
## sample estimates:
##    prop 1    prop 2
## 0.4451877 0.3035422
```

# Observational Data

- It is *extremely unlikely* that male and female applicants to UC-Berkeley are admitted at equal rates... but does that *prove* there is gender-discrimination?

# Observational Data

- It is *extremely unlikely* that male and female applicants to UC-Berkeley are admitted at equal rates... but does that *prove* there is gender-discrimination?
  - No, these data are observational, so there might be other explanations for this association
  - For example, the association might be due to a **confounding variable** that our simple analysis failed to control for

# Observational Data

# UC-Berkeley Takeaways

- It was inappropriate to look at the overall acceptance rates because males and females tended to apply to different departments
  - The overall male rate is boosted by males disproportionately applying to departments A and B, which tend to accept most applicants (regardless of gender)
  - Conversely, females tended to apply to more selective departments that do not accept very many applicants (regardless of their gender)

# UC-Berkeley Takeaways

- It was inappropriate to look at the overall acceptance rates because males and females tended to apply to different departments
  - The overall male rate is boosted by males disproportionately applying to departments A and B, which tend to accept most applicants (regardless of gender)
  - Conversely, females tended to apply to more selective departments that do not accept very many applicants (regardless of their gender)
- Filtering the data by department, a technique known as *stratification*, was essential to figuring this out
  - As you might expect, it becomes difficult to stratify by many variables (we'll revisit this issue when learning about *multiple regression*)

# Infant Heart Surgery

- Roughly 1 in 500 infants are born with congenital heart defects (CHDs) that require surgery shortly after birth
- A study conducted by Harvard Medical school randomly assigned infants born with CHDs to one of two surgical groups

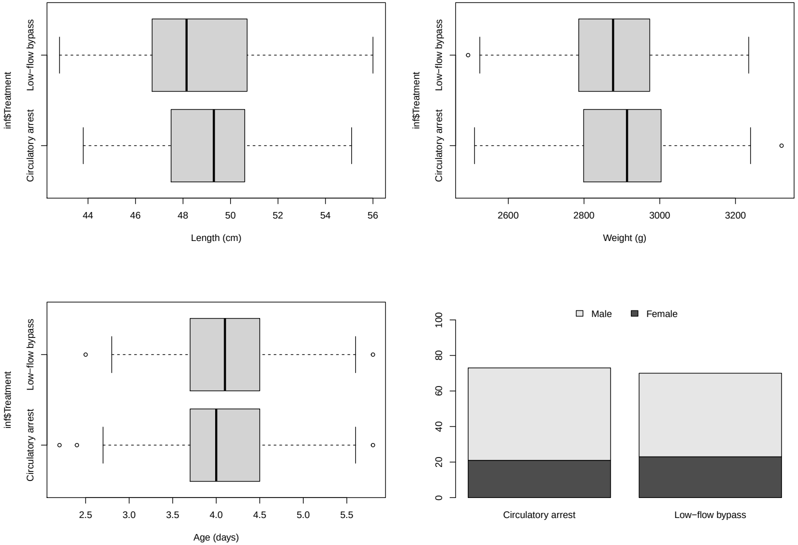
# Infant Heart Surgery

- Roughly 1 in 500 infants are born with congenital heart defects (CHDs) that require surgery shortly after birth
- A study conducted by Harvard Medical school randomly assigned infants born with CHDs to one of two surgical groups
  - *Circulatory Arrest* - the current standard of case that comes with the downside of cutting off the flow of blood to the brain
  - *Low-flow bypass* - a new procedure that uses an external pump to maintain circulation to the brain, but may lead to other types of brain damage

# Infant Heart Surgery

- Roughly 1 in 500 infants are born with congenital heart defects (CHDs) that require surgery shortly after birth
- A study conducted by Harvard Medical school randomly assigned infants born with CHDs to one of two surgical groups
  - *Circulatory Arrest* - the current standard of case that comes with the downside of cutting off the flow of blood to the brain
  - *Low-flow bypass* - a new procedure that uses an external pump to maintain circulation to the brain, but may lead to other types of brain damage
- The researchers compared *psychomotor development* (PDI) and *mental development* (MDI)
  - Infants in the Low-flow group had *significantly higher MDI*. . . but could this be due to a confounding variable?

# The Power of Randomization



# Infant Heart Surgery Takeaways

- When the explanatory variable is randomized, confounding variables are not a concern, as they'll end up being balanced
  - For example, characteristics like height/weight/age/sex were all equally represented in both surgical groups, so they cannot possibly explain the difference in outcomes

# Closing Remarks

- This week, our focus will be on *data exploration*, or the process of identifying relationships in our data using visualization
  - Data visualization is an extremely effect method for identifying *confounding variables*
  - When using `ggplot`, *stratification* is easy to implement using the `facet_wrap` function

# Closing Remarks

- This week, our focus will be on *data exploration*, or the process of identifying relationships in our data using visualization
  - Data visualization is an extremely effect method for identifying *confounding variables*
  - When using `ggplot`, *stratification* is easy to implement using the `facet_wrap` function
- Even when the data come from a randomized experiment, visualization provides an effect means for checking that the randomization was properly executed
  - And, as we'll see on Thursday, data visualization can guide us through *data transformations* that can make our models more effective