Lecture 19

Limitations of randomized experiments

Julian Reif
Fall 2025

RStudio setup for this lecture

- Log into RStudio on your Amazon EC2 instance
  - Use AMI `FIN550-RStudio` with IAM role `BigDataEC2Role`

```
# This is a Unix command. Enter via RStudio Terminal
aws s3 cp --recursive s3://bigdata-fin550-reif/lecture-19 ~/fin550/lecture-19
```

Randomized experiments are not perfect

- Randomized experiments are a powerful tool for using data to answer causal questions

- Nevertheless, randomized experiments still have limitations

- Today:
  1. Examine those limitations
  2. Learn how to estimate causal effects of randomized experiments in R

Common pitfalls and limitations

1. Randomization failure (selection bias)

2. Attrition bias

3. External validity

4. Statistical power

Limitation 1: randomization failure

Randomization failure

- If done correctly, randomization removes selection bias

- But, randomization can fail for several reasons:
  1. By chance, treatment and control groups are unbalanced
  2. Researcher fails to properly assign groups
  3. Participants refuse to comply with their treatment assignments

- Checking for balance and verifying treatment helps mitigate these concerns

Example: effect of high school tutoring program

- Researcher randomly assigns high school students to a tutoring program
  - Forms a partnership with a local high school
  - Teachers are responsible for randomly assigning students to tutoring program

- Suppose some teachers preferentially assign struggling students to tutoring program

- What will happen to the estimated causal treatment effect?
  - Selection bias

Limitation 2: attrition bias

Attrition happens in many randomized experiments

- Experimental studies typically include data only on individuals who complete the study

- **Attrition** occurs when an individual leaves a study before completion
  - Death
  - Not willing to continue participation
  - No longer available
  - Geographical move
  - Adverse response to treatment

- Bias can arise when attrition occurs *non-randomly*
  - Can the control group still serve as a valid comparison for the treatment group?

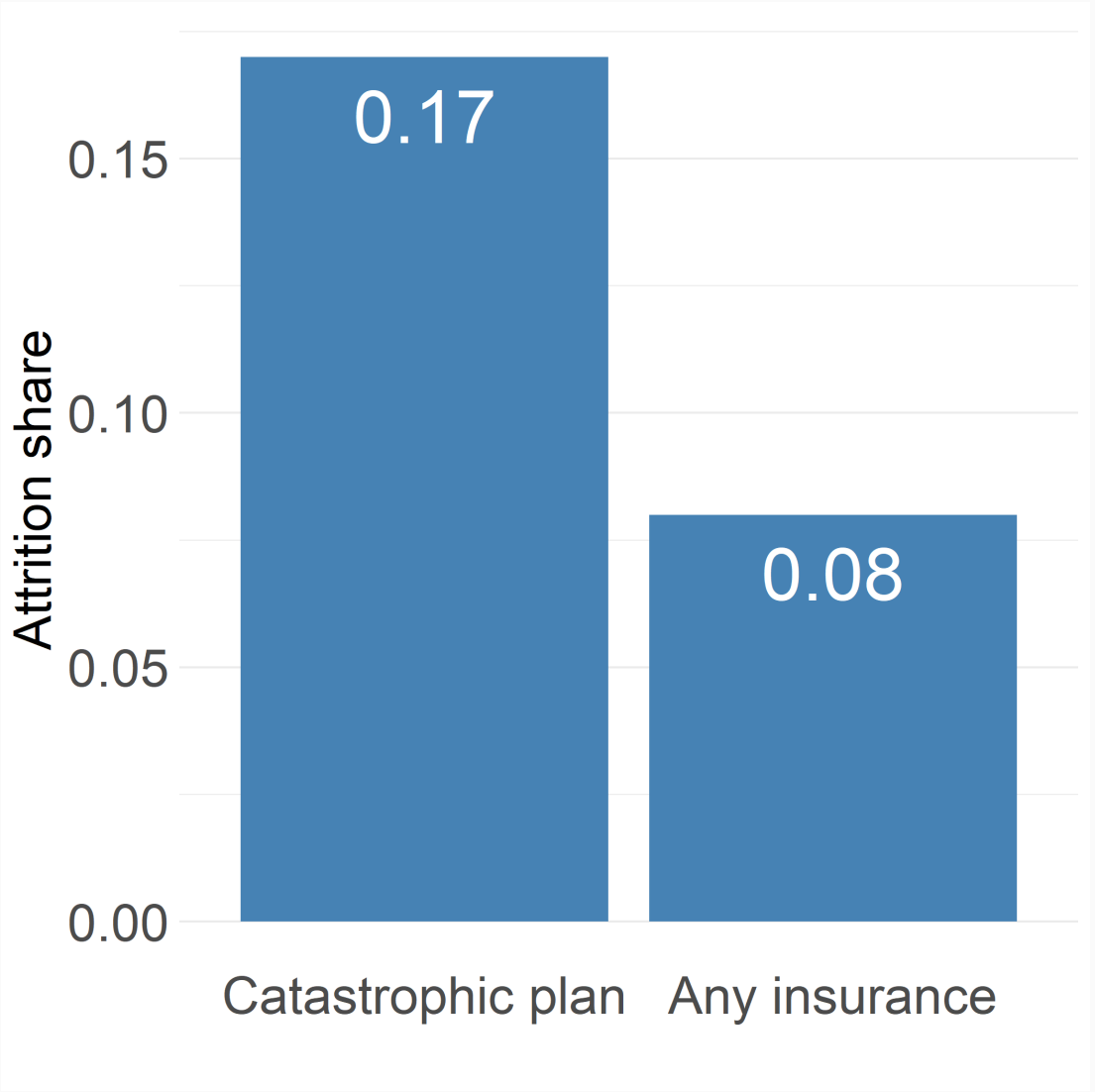
Example: RAND Health Insurance Experiment

- Random assignment to health insurance plans
  - Control group: catastrophic plan (minimal insurance)
  - Treatment group: all other plans ("any insurance")

- Can you force people to stay in your study? No!

- Who do you think is more likely to leave the RAND study:
  - Somebody randomly assigned to free/generous health insurance?
  - Or somebody assigned to the catastrophic plan?

Attrition shares in RAND Health Insurance Experiment



Attrition bias in RAND Health Insurance Experiment

- Differential attrition may cause the control group to differ from the treatment group
  - That can lead to selection bias!

- Why treatment and control groups might differ in RAND:
  - Few people leave treatment group ("any insurance"), where healthcare is cheap
  - In control group, people expecting high medical expenses are most likely to leave
  - Thus, more high-spending people end up in the free care plan
  - This attrition bias makes it look like free care leads to higher spending!

- Recently, researchers have tried to adjust for this bias
  - Overall conclusion is that the bias is small

Limitation 3: external validity

Randomized experiments may not be generalizable

- **Internally valid** estimates describe the true causal effect among subjects
  - That is, the estimate does not suffer from selection bias

- **Externally valid** estimates also describe the causal effect among non-subjects

- External validity is questionable when:
  - Experimental subjects differ from the non-experimental subjects
  - Experimental treatment differs from real-world treatments

Example: external validity of a tutoring program

- Research experiment finds that University of Illinois tutoring program improves learning

- Can we conclude that University of Michigan's tutoring program also improves learning?
  - Maybe not, if the tutoring programs are different

- Can we conclude that tutoring programs for high school students improve learning?
  - Maybe not, if we think high school students respond differently to tutoring programs

External validity of RAND Health Insurance Experiment

- How generalizable are the results from RAND?
- Can we conclude that health insurance increases medical spending *in 2022?*
- Can we conclude that health insurance increases medical spending *in Brazil?*
- Can we conclude that *dental* insurance increases *dental* spending?

Limitation 4: statistical power

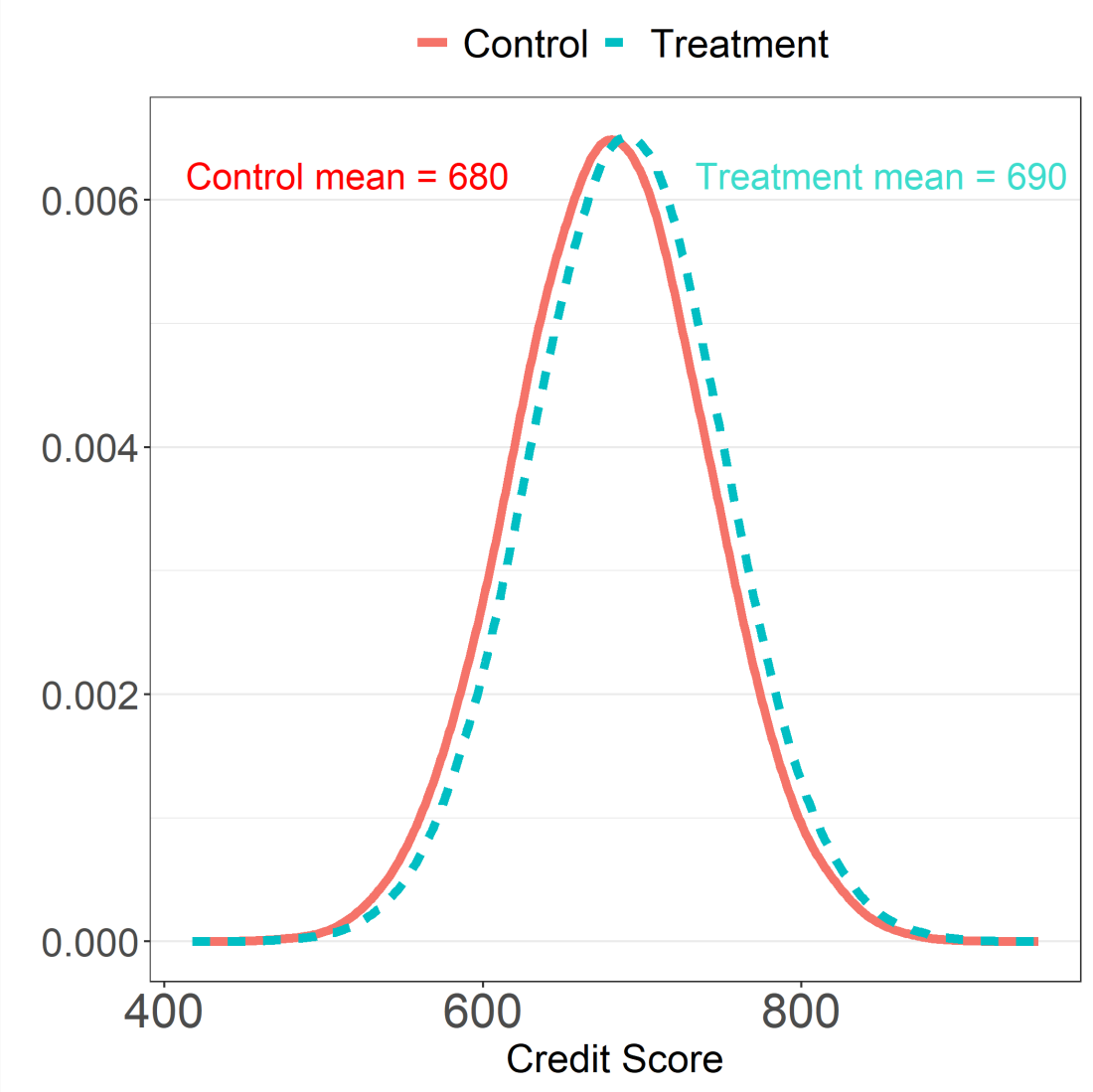
Estimates vary due to randomness in the data

- The larger the sample, the more precise the estimate

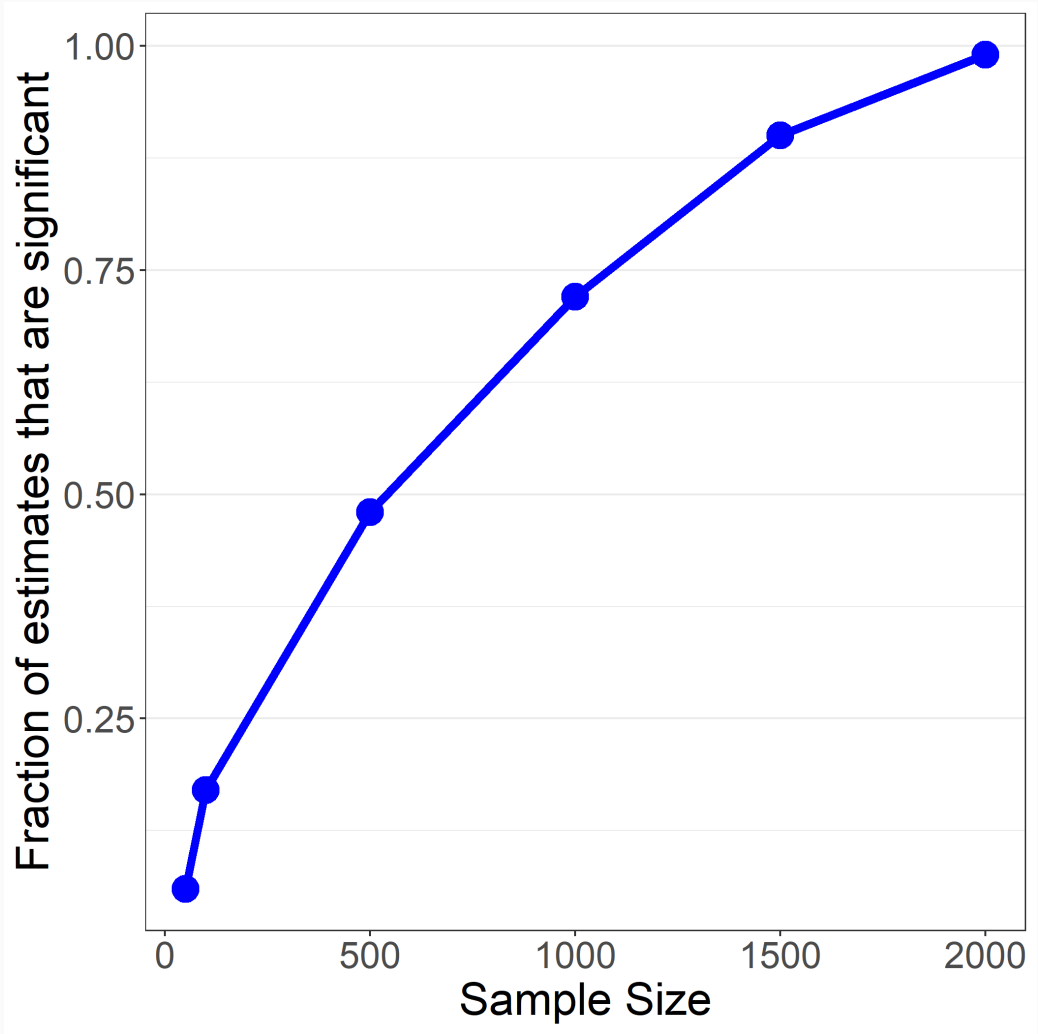
- "Statistical power" refers to ability to detect the causal effect

- Randomized experiments with small sample sizes have low statistical power
  - Confidence intervals are large
  - The causal effect may be lost in the noise

Example: effect of financial education on credit scores



How often do we estimate a statistically significant effect?



- Regression:
  $SCORE = \beta_0 + \beta_1 TREAT + \epsilon$
- For each sample size, generate 100 random datasets
  - Estimate regression for each dataset
- What fraction of estimates are statistically significant?

Randomized experiments in R

Estimating the effect of health insurance on spending

Consider the following data generating process:

$$SPENDING_i = \beta_0 + \beta_1 TREAT_i + \beta_2 AGE_i + \epsilon_i$$

where

- $SPENDING_i$ is healthcare spending (thousands of dollars) by person i
- $TREAT_i \in \{0, 1\}$ indicates whether person i has randomly assigned health insurance
- $AGE_i \in [18, 100]$ is age of person i
- ϵ_i is a mean-zero random error

We are interested in estimating β_1: the causal effect of health insurance on spending

Function to create a simulated dataset

```r
library(tidyverse)
library(ggplot2)
library(broom)

# Function to create dataset with N observations
data_sample <- function(N = 100, beta0 = 10, beta1 = 1, beta2 = 0.1, sd_e = 3) {
  data <- tibble(
    age   = sample(seq(18,100), N, replace=T),
    treat = sample(c(0,1), N, replace=T),
    spending = beta0 + beta1*treat + beta2*age + rnorm(N, mean = 0, sd = sd_e))

  data <- data %>% relocate(spending, .before = age)

  return(data)
}
```

Create and inspect dataset

```
set.seed(10)
df <- data_sample(N=100)

summary(df$spending)
cat("\n")
head(df)
```

```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#   7.372  14.154  16.926  17.059  18.843  29.688
#
# # A tibble: 6 × 3
#   spending   age treat
#      <dbl> <int> <dbl>
# 1     15.5    26     0
# 2     18.5    91     1
# 3     18.6    93     0
# 4     16.4    72     0
# 5     19.7    89     1
# 6     21.6    71     1
```

Try it: balance test for age

```
# Calculate mean of age, for control and treatment groups




# Calculate the difference in means for age (treat - control)
```

Balance test for age

```r
# Calculate mean of age, for control and treatment groups
df %>%
  group_by(treat) %>%
  summarize(mean_age = mean(age))

# Calculate the difference in means for age (treat - control)
mean(df$age[df$treat==1]) - mean(df$age[df$treat==0])
```

```
# # A tibble: 2 × 2
#   treat mean_age
#   <dbl>    <dbl>
# 1     0     57.0
# 2     1     60.4
# [1] 3.401044
```

Balance test for age: regression

```
# Is the difference in age means statistically significant?
lm1 <- lm(age ~ treat, data = df)
tidy(lm1, conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 7
#   term        estimate std.error statistic  p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
# 1 (Intercept)    57.0       3.49    16.3   1.00e-29    50.0       63.9
# 2 treat           3.40      4.79     0.710 4.79e- 1    -6.10      12.9
```

Try it: estimate the causal treatment effect

```
# Calculate difference in means for spending (treat - control)



# Use regression to estimate whether the difference statistically significant
```

Estimate the causal treatment effect

```r
# Calculate difference in means for spending (treat - control)
mean(df$spending[df$treat==1]) - mean(df$spending[df$treat==0])
cat("\n")

# Use regression to estimate whether the difference statistically significant
lm2 <- lm(spending ~ treat, data=df)
tidy(lm2, conf.int = T, conf.level = 0.95)
```

```
# [1] 1.490277
#
# # A tibble: 2 × 7
#   term        estimate std.error statistic  p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
# 1 (Intercept)    16.3      0.613     26.6  1.44e-46   15.1       17.5
# 2 treat           1.49     0.841      1.77 7.96e- 2   -0.179      3.16
```

Try again with a larger sample size

```
# With N=1000, estimate is more precise
set.seed(10)
df2 <- data_sample(N=1000)
lm3 <- lm(spending ~ treat, data=df2)
tidy(lm3, conf.int = T, conf.level = 0.95)
```

```
# # A tibble: 2 × 7
#   term        estimate std.error statistic    p.value conf.low conf.high
#   <chr>          <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
# 1 (Intercept)    15.9      0.174     91.2  0             15.5       16.2
# 2 treat           1.13     0.250      4.53 0.00000664    0.642      1.62
```

Summary

- Randomized experiments are powerful but still have limitations
  - Randomization failure and attrition bias threaten internal validity
  - External validity is always a concern

- If sample sizes are small, experiments may not have enough power to detect effects

- Linear regression can be used to calculate difference in means
  - Regression also provides standard errors, p-values, confidence intervals, etc.