Week 3, Class 6: Practice Exercises - ANSWER KEY

Research Designs

2024-12-31

1 Non-AI Exercises

1.1 1. Understanding Research Designs

1.1.1 1.1 Match: Research Design Types

Match each research design with its key characteristic:

Designs: a) Experimental b) Natural experiment c) Cross-sectional d) Panel/Longitudinal

Characteristics: 1. Observes the same units over multiple time periods 2. Researcher controls random assignment 3. Takes advantage of external events for quasi-random assignment 4. Observes many units at a single point in time

Matches: a = **2**, b = **3**, c = **4**, d = **1**

1.1.2 1.2 Multiple Choice: Random Assignment

What is the main advantage of random assignment in experiments?

a) It makes the study cheaper to conduct
b) It ensures treatment and control groups are similar on average
c) It guarantees everyone benefits from treatment
d) It eliminates the need for statistical analysis

Answer: **b) It ensures treatment and control groups are similar on average**

1.1.3 1.3 True or False: Research Designs

Mark each statement as True (T) or False (F):

F Experiments always require a laboratory setting **T** Natural experiments rely on events outside researcher control **F** Cross-sectional data can show individual change over time **T** Panel data follows the same people over multiple periods **F** Observational studies can never establish causation

1.2 2. Natural Experiments

1.2.1 2.1 Fill in the Blanks: Natural Experiments

Natural experiments occur when:

1. An external **event** creates quasi-random assignment
2. The researcher does not **control** the treatment
3. Groups become comparable by **chance**
4. We can compare **treatment** and control groups
5. The assignment process is **unrelated** to political outcomes

1.2.2 2.2 Code Detective: Research Design

What type of research design does this analysis suggest?

```
data %>%
  filter(distance_to_border < 10) %>%
  mutate(treatment = ifelse(state == "Legal_Marijuana", 1, 0)) %>%
  group_by(treatment) %>%
  summarise(avg_crime = mean(crime_rate))
```

This code suggests a: **Natural experiment using geographic borders (border discontinuity design)**

1.2.3 2.3 Multiple Choice: Lottery Draft

The Vietnam draft lottery is a good natural experiment because:

a) Researchers controlled who was drafted
b) Birth dates were randomly assigned to draft numbers
c) Everyone wanted to avoid the draft

d) It only affected men

Answer: **b) Birth dates were randomly assigned to draft numbers**

1.3 3. Cross-sectional vs Longitudinal Data

1.3.1 3.1 Match: Data Structure

Match each scenario with the appropriate data type:

Scenarios: a) Survey 1000 voters on election day b) Track 500 families' income for 10 years c) Poll different people each month about approval d) Interview same legislators every session

Data Types: 1. Cross-sectional 2. Panel/Longitudinal 3. Repeated cross-sections 4. Time series panel

Matches: a = **1**, b = **2**, c = **3**, d = **2** (or 4 if focusing on time series aspect)

1.3.2 3.2 Fill in the Code: Panel Data

Complete this code to analyze panel data:

```
panel_data %>%
  group_by(respondent_id) %>%  # Group by individual
  arrange(wave) %>%   # Sort by time
  mutate(
    income_change = income - lag(income)
  )
```

1.3.3 3.3 Spot the Error: Research Design

What's wrong with this conclusion?

"We surveyed 1000 people and found that older people vote more. Therefore, as people age, they become more likely to vote."

Problem: **This uses cross-sectional data (different people of different ages at one time) to make a claim about change over time. We can't tell if it's an age effect or a cohort/generation effect. We would need panel data following the same people over time.**

1.4 4. Validity and Limitations

1.4.1 4.1 Multiple Choice: Internal Validity

A study has high internal validity when:

a) Results apply to many different contexts
b) We can confidently attribute effects to the treatment
c) The sample size is very large
d) It uses advanced statistical methods

Answer: **b) We can confidently attribute effects to the treatment**

1.4.2 4.2 Match: Validity Threats

Match each threat with its type:

Threats: a) Results only apply to college students b) Something else caused the outcome c) People dropped out of the study d) Treatment and control groups were different

Types: 1. Selection bias 2. External validity 3. Attrition 4. Confounding

Matches: a = **2**, b = **4**, c = **3**, d = **1**

1.4.3 4.3 True or False: Research Tradeoffs

Mark each statement as True (T) or False (F):

T Lab experiments have high internal validity but may lack realism **F** Natural experiments always have perfect random assignment **F** Large samples guarantee causal inference **T** Field experiments balance control and realism **F** Observational studies are always inferior to experiments

2 AI Exercises

2.1 5. Analyzing a Natural Experiment

Dataset: border_policy_change.csv

2.1.1 5.1 Understanding the Natural Experiment

```r
# Load the dataset
library(tidyverse)
border_data <- read_csv("border_policy_change.csv")

# Examine the structure
glimpse(border_data)
```

```
Rows: 3,500
Columns: 12
$ county_fips       <dbl> 53506, 43297, 58596, 34619, 43094, 37427, 53506, 491~
$ year              <dbl> 2015, 2020, 2017, 2020, 2016, 2012, 2010, 2020, 2010~
$ unemployment_rate <dbl> 4.2, 8.7, 7.3, 4.9, 2.4, 12.2, 3.3, 2.4, 12.4, 11.4,~
$ median_income     <dbl> 63262, 61006, 48323, 55327, 66234, 57395, 55070, 462~
$ gini_index        <dbl> 0.35, 0.49, 0.36, 0.54, 0.38, 0.51, 0.40, 0.54, 0.45~
$ poverty_rate      <dbl> 15.1, 24.1, 26.0, 5.7, 30.0, 25.6, 25.3, 22.5, 27.4,~
$ pop_density       <dbl> 201.3, 229.9, 123.7, 372.2, 273.8, 522.0, 266.8, 94.~
$ urban_rural       <chr> "Urban", "Rural", "Rural", "Urban", "Urban", "Suburb~
$ percent_white     <dbl> 84.4, 79.1, 21.3, 34.1, 73.2, 71.7, 68.8, 41.1, 41.2~
$ percent_black     <dbl> 3.1, 56.3, 32.7, 3.6, 19.5, 33.3, 42.8, 5.9, 51.5, 1~
$ percent_hispanic  <dbl> 23.7, 17.8, 49.2, 56.1, 47.3, 10.4, 32.7, 52.9, 14.9~
$ wage_change       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRU~
```

2.1.2 5.2 Implementing the Design

```r
# Compute poverty rates and unemployment in treated and
# control counties
results <- border_data %>%
    group_by(wage_change) %>%
    summarise(mean_poverty = mean(poverty_rate, na.rm = TRUE),
        mean_unemployment = mean(unemployment_rate, na.rm = TRUE),
        mean_income = mean(median_income, na.rm = TRUE), n = n())

print(results)
```

```
# A tibble: 2 x 5
  wage_change mean_poverty mean_unemployment mean_income     n
  <lgl>              <dbl>             <dbl>       <dbl> <int>
```

```
1 FALSE                17.5              8.36      52067.  2517
2 TRUE                 17.5              8.66      51623.   983
```

```
# Calculate differences
treated_poverty <- results$mean_poverty[results$wage_change ==
    TRUE]
control_poverty <- results$mean_poverty[results$wage_change ==
    FALSE]
diff_poverty <- treated_poverty - control_poverty

treated_unemployment <- results$mean_unemployment[results$wage_change ==
    TRUE]
control_unemployment <- results$mean_unemployment[results$wage_change ==
    FALSE]
diff_unemployment <- treated_unemployment - control_unemployment

print(paste("Poverty difference:", round(diff_poverty, 2), "percentage points"))
```

```
[1] "Poverty difference: -0.03 percentage points"
```

```
print(paste("Unemployment difference:", round(diff_unemployment,
    2), "percentage points"))
```

```
[1] "Unemployment difference: 0.3 percentage points"
```

2.2 6. Working with Panel Data

Dataset: voter_panel_study.csv

2.2.1 6.1 Exploring Panel Structure

```
# Load the dataset
panel_data <- read_csv("voter_panel_study.csv")

# Check the panel structure
glimpse(panel_data)
```

```
Rows: 3,000
Columns: 10
$ respondent_id         <dbl> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6~
$ wave                  <dbl> 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2~
$ age                   <dbl> 78, 78, 78, 74, 74, 74, 61, 61, 61, 33, 33, 33, 7~
$ education             <dbl> 4, 4, 4, 4, 4, 4, 5, 5, 5, 2, 2, 2, 5, 5, 5, 3, 3~
$ ideology              <dbl> 2, 2, 2, 2, 2, 2, 3, 3, 3, 7, 7, 7, 4, 4, 4, 7, 7~
$ political_interest    <dbl> 2, 7, 5, 3, 6, 5, 6, 5, 3, 3, 5, 4, 3, 6, 5, 6, 4~
$ social_media_use      <dbl> 5, 1, 1, 3, 3, 0, 3, 3, 5, 4, 2, 5, 4, 4, 2, 2, 3~
$ participated_protest <dbl> 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
$ protest_issue         <chr> "Climate", NA, NA, "Racial Justice", "Economy", "~
$ voted_last_election  <dbl> 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1~
```

```r
# How many waves and respondents?
n_waves <- panel_data %>%
    summarise(n_waves = n_distinct(wave))
print(paste("Number of waves:", n_waves$n_waves))
```

```
[1] "Number of waves: 3"
```

```r
n_respondents <- panel_data %>%
    summarise(n_resp = n_distinct(respondent_id))
print(paste("Number of respondents:", n_respondents$n_resp))
```

```
[1] "Number of respondents: 1000"
```

2.2.2 6.2 Analyzing Change Over Time

```r
# Track how individual respondents' ideology and political interest change
change_data <- panel_data %>%
  arrange(respondent_id, wave) %>%
  group_by(respondent_id) %>%
  mutate(
    # Calculate change from previous wave
    interest_change = political_interest - lag(political_interest),
    ideology_change = ideology != lag(ideology)
  )

# Summary of changes
```

```r
change_summary <- change_data %>%
  filter(!is.na(interest_change)) %>%
  summarise(
    mean_interest_change = mean(interest_change, na.rm = TRUE),
    pct_ideology_change = mean(ideology_change, na.rm = TRUE) * 100
  )

print("Average change in political interest between waves:")
```

```
[1] "Average change in political interest between waves:"
```

```r
print(change_summary$mean_interest_change)
```

```
  [1]  1.5  1.0 -1.5  0.5  1.0 -0.5 -2.0 -0.5  0.0  1.0  1.0  2.0  2.5  3.0
 [15]  0.0 -1.0  1.5 -1.0 -1.0  0.0 -0.5 -2.0  0.0  0.0 -2.0 -1.0  1.5  1.5
 [29] -2.0 -1.0  0.0  0.5  2.0 -3.0 -1.0 -0.5 -1.5 -1.5 -1.5  0.0 -1.0 -2.0
 [43]  0.0  1.0  1.0  0.0  0.5 -4.0 -0.5 -1.5  0.5 -1.5  0.5  0.5 -1.0  0.5
 [57] -0.5  1.0  0.5  0.5  0.5 -1.5  1.0  0.0  2.5  1.5  0.5 -1.0  0.5  0.0
 [71]  1.5  0.5  1.5 -1.0  0.0  1.5  0.0 -0.5 -0.5  2.0 -1.5 -2.0  0.0  0.5
 [85]  0.5  1.5 -1.0  0.0  0.5  2.5 -0.5  0.5  2.0 -4.0  0.5  0.0  0.0  2.0
 [99] -2.0  1.0  1.0  1.0 -3.0 -0.5 -1.0  0.5  2.0  0.0  2.5 -1.0 -0.5 -1.5
[113]  0.5 -3.0 -1.5  0.5  0.0  0.0  0.5 -0.5 -0.5  2.5  1.5 -1.0 -2.5  2.0
[127] -0.5  1.5  1.5 -2.0  0.5  0.5  0.0  0.0 -1.0  0.5  0.5 -1.0 -2.0 -1.5
[141]  1.5  3.5 -2.0  0.5 -0.5  1.0  1.5  1.0 -2.0 -1.0 -0.5  1.0  0.5  1.5
[155]  0.5  0.0  2.5 -0.5  3.0  1.0  1.5  0.0 -0.5  1.5  1.5 -1.0 -1.0  0.0
[169]  0.0  0.0 -1.0  1.0  1.0  0.5  1.5 -0.5 -0.5 -0.5  0.0 -2.5  0.5  0.5
[183]  1.0 -2.0  2.5  0.5  1.5  1.5  0.5  0.0  1.5 -1.0  0.5 -3.0  0.0  1.0
[197]  0.0  1.0  0.0 -0.5 -0.5 -1.5  0.5  1.5  1.5  1.5 -3.0 -2.0  0.0  1.0
[211]  0.5  0.5  1.0  0.0  0.0 -1.0 -0.5 -0.5  1.5  0.0  0.0 -2.5  1.0  0.0
[225] -0.5 -0.5 -3.0  2.0 -1.5  0.5  0.0  0.5  2.0 -1.0  0.5  3.0 -1.0  1.5
[239]  0.0  0.0  0.5  0.0  1.0 -3.0  2.0  1.5 -1.5  2.0  0.0 -1.0  0.5  0.0
[253]  2.0  0.5  0.0 -1.0  0.5  0.0 -1.5  1.5  1.5 -0.5 -2.0  0.5  0.0 -0.5
[267]  0.0 -0.5  1.0 -1.5 -2.5 -1.0 -1.5  0.0 -1.0 -0.5 -1.0  0.0 -0.5 -2.0
[281]  0.5 -1.0 -1.5  0.5  1.5  1.0 -0.5  0.5 -1.0 -1.5  2.5 -1.0  0.5  0.0
[295]  0.5 -2.5  3.5  0.5  0.0 -1.5 -1.0  0.0 -1.0  1.0  1.5 -1.0 -1.5  0.5
[309]  2.0 -0.5  0.5 -3.0 -0.5  1.0  1.0  1.5 -2.0  1.5  0.0 -0.5 -0.5  0.0
[323]  1.0 -2.0 -0.5 -1.5  1.5  1.5  0.0  1.5  2.0 -1.5 -1.0 -0.5  2.0 -1.0
[337] -1.0 -2.0  1.5  0.5  0.0  1.5 -1.0  1.0  0.5  3.0  0.5  1.0 -1.0  0.5
[351] -0.5 -1.0  0.0  1.5  0.0  1.0 -0.5  0.0  0.0  0.0  3.5  0.5 -2.5  1.0
[365]  1.0 -0.5  1.0 -2.0  0.0 -0.5 -1.5  2.0  1.0 -0.5  0.0  0.5  1.5  1.0
[379]  0.5 -1.0 -1.0 -0.5 -2.5  1.0 -0.5 -2.5  0.5 -1.5 -2.0  1.5 -0.5 -2.5
```

```
[393]  0.0  0.5 -1.5 -1.5  1.5 -1.0  0.0  1.5  0.5  0.0 -2.0 -1.0  1.0 -1.0
[407]  0.5  2.0  0.5  0.0 -1.0  0.0 -2.0 -1.0  1.0 -2.5 -1.0 -0.5 -1.5 -3.0
[421] -0.5 -1.0  0.0  1.0 -2.5  0.5  1.0 -2.0 -1.0 -1.0  3.0  0.5  2.0 -1.0
[435] -0.5 -1.5  0.0 -1.5  1.0  0.5  0.0  0.5 -0.5 -2.0 -2.0 -0.5 -0.5  2.0
[449]  2.0  0.5  1.0  0.0  1.5 -2.0  2.0  2.0  0.5  0.0 -1.0 -0.5  1.0 -0.5
[463]  0.5  0.5  2.5 -0.5 -2.0  2.0 -2.5 -0.5 -1.0  1.0 -0.5 -4.0  2.5  0.5
[477]  0.0 -2.5  0.5  2.0  1.0  1.0 -1.0  0.5 -1.0 -2.0  0.5  0.0 -2.5 -1.0
[491] -0.5 -1.5 -1.0 -0.5 -3.0  1.5  0.0 -1.5  2.5 -0.5  0.0  1.0  0.5 -0.5
[505]  0.5 -2.5 -2.0 -1.0 -2.0 -2.5  2.5 -1.5  0.0  1.5  0.5 -1.0  1.0  2.5
[519]  2.0 -0.5  0.0 -0.5 -0.5  0.5  0.0  1.5  1.5 -0.5  0.0 -2.0 -1.5  0.0
[533] -0.5  0.0 -1.0 -1.5 -1.0  1.0  1.5 -2.5 -4.0  2.5 -0.5  1.0  1.0  1.0
[547]  2.5 -2.5 -2.0 -1.5  1.5 -2.0  0.5 -2.5  2.5  0.0  1.0  0.5 -0.5  0.5
[561]  0.0  1.5  0.5 -2.0  0.5  0.5 -0.5  0.5  0.5  1.0 -1.5 -1.0 -1.0 -1.5
[575]  1.5  0.5 -1.5  1.0 -2.5 -0.5  1.5  1.5  0.0  1.5 -1.0  0.0  0.5 -1.0
[589]  0.5  0.5 -3.5 -1.5 -1.0  1.0  2.5 -0.5 -1.0 -1.5  2.0  0.5  0.0  1.5
[603]  1.5  0.0 -2.5 -1.0 -0.5 -1.5  0.0  0.0  0.5  0.0  1.0 -0.5  0.5 -1.5
[617]  1.0  2.0  0.0 -0.5  2.0  1.5 -1.0 -1.5  0.5  2.5 -0.5  2.0  0.0  0.5
[631] -1.5  1.5  3.0 -1.5 -0.5  0.5 -2.0 -0.5  3.0  0.5  1.0 -0.5  1.0  1.0
[645] -2.5 -1.0 -0.5  0.0 -1.0 -1.0 -0.5 -0.5 -3.0 -1.0  3.0  1.0  2.5 -0.5
[659]  0.5 -0.5 -1.5 -1.5  1.5 -1.5  0.0  1.0  0.0 -1.0 -3.5 -0.5  2.5 -1.5
[673]  1.5  0.0  2.0  1.0 -0.5  0.0  2.5  1.0 -1.0  2.0  1.0  1.0  1.0  0.0
[687] -0.5 -0.5 -1.5  1.0  1.0 -0.5  1.0  0.5  0.0  1.5  0.5  3.0  2.0  1.0
[701]  0.5  0.5 -0.5  0.5  3.5  1.0  2.5  1.5 -1.0 -1.5 -1.5  2.5 -0.5  0.0
[715] -3.5  1.0 -0.5 -1.5  1.0 -3.0  2.0 -1.0  1.5 -0.5 -1.0  3.0  0.5  1.5
[729]  0.5 -1.0 -2.0 -0.5  0.0  2.0  0.5 -0.5 -0.5  1.0  0.5  1.0  1.5  0.0
[743]  0.5  0.5  0.0 -0.5  1.5  0.5  2.0  0.0 -2.0  2.0  0.0 -0.5 -1.0  1.5
[757]  1.0 -0.5 -1.0  1.5  1.0  0.0  0.0  0.0 -1.5 -1.0  1.0  0.5 -1.0 -2.0
[771]  3.0 -2.0 -3.0 -2.0 -1.0  2.0 -2.0  1.5  0.0 -1.5  0.0  0.5  1.5  0.5
[785]  2.5  0.0  1.5  0.5  0.0 -0.5 -2.5  1.0  0.0  0.5  2.0  1.0  2.5 -2.0
[799] -2.0 -0.5 -0.5  1.0 -1.0  0.0 -2.5 -2.0 -1.5  0.0  1.5 -1.0 -0.5  2.5
[813]  1.5  1.5 -3.0 -0.5 -0.5  0.0  0.0 -1.0 -4.0  2.0  2.0 -1.5 -1.5 -0.5
[827] -2.0 -0.5  1.0  1.5  1.0  1.0  0.5  1.0  0.5  2.0 -1.0  0.0  0.5 -1.0
[841] -2.0  0.5  1.0  1.0  0.0 -2.0 -2.0  3.5  2.0  1.5  2.5  2.5  0.0 -0.5
[855]  0.0  1.0 -3.0  0.0 -1.0 -2.5 -0.5  1.5 -2.5 -0.5  0.0 -1.0  3.0 -3.5
[869] -3.0 -2.0 -1.0 -0.5 -0.5 -0.5  3.5 -3.5  0.5 -1.0 -0.5  1.0 -3.0  0.5
[883] -1.0  1.0  1.5  1.0  1.5  0.0  1.0 -2.5  1.5  0.0  0.0  1.0 -0.5 -1.5
[897] -1.0 -1.5 -1.0 -0.5 -2.5 -0.5  1.0  1.0  0.0  1.5  1.5 -0.5 -1.0  0.5
[911] -0.5  2.5 -1.5 -2.0 -2.0  1.0 -2.5 -1.5 -0.5 -2.0 -2.0 -2.0 -2.0  0.5
[925]  0.0 -3.5 -1.0 -0.5 -1.5  0.0  1.0  1.0  0.5  1.5 -1.0  0.5 -0.5  0.5
[939]  1.0 -1.0  3.0  0.0 -1.0 -0.5  1.5 -1.0 -2.5  1.0 -2.5  1.0  1.0  1.5
[953] -2.0  0.0  1.5  1.5 -1.0  0.5 -1.5 -0.5  1.0 -0.5  1.0 -1.0  0.0  1.0
[967] -2.0 -1.0  1.5  0.0  0.5 -0.5 -0.5  0.0  0.0 -0.5 -0.5 -2.0 -0.5  0.5
[981] -3.0  2.0  2.5 -0.5 -1.0  1.5  0.0  0.5  1.5 -1.5 -3.0  2.0 -2.0  1.0
```

```
 [995]  1.0  1.5 -1.5  0.0  0.0 -0.5
```

```r
print("Percent who changed ideology between waves:")
```

```
[1] "Percent who changed ideology between waves:"
```

```r
print(change_summary$pct_ideology_change)
```

```
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [260] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [297] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [667] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [704] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [741] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [778] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [815] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [889] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [926] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[1000] 0
```

```r
# Is average social media higher or lower for those who voted?
social_media_voting <- panel_data %>%
  group_by(voted_last_election) %>%
  summarise(
```

```r
    mean_social_media = mean(social_media_use == "Daily", na.rm = TRUE) * 100,
    n = n()
  )

print("Social media use by voting behavior:")
```

```
[1] "Social media use by voting behavior:"
```

```r
print(social_media_voting)
```

```
# A tibble: 2 x 3
  voted_last_election mean_social_media     n
                <dbl>             <dbl> <int>
1                   0                 0   890
2                   1                 0  2110
```

```r
# Identify voters who become more/less engaged
engagement_changes <- panel_data %>%
  arrange(respondent_id, wave) %>%
  group_by(respondent_id) %>%
  summarise(
    first_voted = first(voted_last_election),
    last_voted = last(voted_last_election),
    engagement_change = case_when(
      first_voted == FALSE & last_voted == TRUE ~ "Became engaged",
      first_voted == TRUE & last_voted == FALSE ~ "Became disengaged",
      first_voted == last_voted ~ "No change",
      TRUE ~ "Unknown"
    )
  )

engagement_summary <- engagement_changes %>%
  count(engagement_change) %>%
  mutate(percent = n / sum(n) * 100)

print("Changes in voter engagement:")
```

```
[1] "Changes in voter engagement:"
```

```r
print(engagement_summary)
```

```
# A tibble: 3 x 3
  engagement_change     n percent
  <chr>             <int>   <dbl>
1 Became disengaged   209    20.9
2 Became engaged      212    21.2
3 No change           579    57.9
```

2.3 7. Experimental Design Analysis

Dataset: gotv_experiment.csv

2.3.1 7.1 Understanding the Experiment

```r
# Load the dataset
gotv <- read_csv("gotv_experiment.csv")

# Explore the experimental design
glimpse(gotv)
```

```
Rows: 5,000
Columns: 5
$ voter_id    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ treatment   <chr> "Phone Call", "Phone Call", "Postcard", "Postcard", "Postca~
$ age_group   <chr> "50-64", "50-64", "50-64", "50-64", "30-49", "18-29", "65+"~
$ education   <chr> "High School", "Some College", "High School", "Graduate", "~
$ voted_2022 <dbl> 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1,~
```

```r
# Check treatment groups
gotv %>%
    count(treatment) %>%
    mutate(percent = n/sum(n) * 100)
```

```
# A tibble: 4 x 3
  treatment          n percent
  <chr>          <int>   <dbl>
1 Control         1241    24.8
```

```
2 Personal Visit  1267    25.3
3 Phone Call      1254    25.1
4 Postcard        1238    24.8
```

2.3.2 7.2 Analyzing Treatment Effects

```r
# Calculate turnout rates by treatment group
turnout_by_treatment <- gotv %>%
    group_by(treatment) %>%
    summarise(turnout_rate = mean(voted_2022, na.rm = TRUE) *
        100, n = n()) %>%
    arrange(desc(turnout_rate))

print("Turnout rates by treatment:")
```

```
[1] "Turnout rates by treatment:"
```

```r
print(turnout_by_treatment)
```

```
# A tibble: 4 x 3
  treatment      turnout_rate     n
  <chr>                 <dbl> <int>
1 Personal Visit         68.7  1267
2 Phone Call             63.2  1254
3 Postcard               62.4  1238
4 Control                56.4  1241
```

```r
# Get control group turnout
control_rate <- turnout_by_treatment$turnout_rate[turnout_by_treatment$treatment ==
    "Control"]

# Compare each treatment to control
treatment_effects <- turnout_by_treatment %>%
    filter(treatment != "Control") %>%
    mutate(effect_vs_control = turnout_rate - control_rate) %>%
    arrange(desc(effect_vs_control))

print("Treatment effects compared to control:")
```

```
[1] "Treatment effects compared to control:"
```

```r
print(treatment_effects)
```

```
# A tibble: 3 x 4
  treatment      turnout_rate     n effect_vs_control
  <chr>                 <dbl> <int>             <dbl>
1 Personal Visit         68.7  1267             12.3
2 Phone Call             63.2  1254              6.83
3 Postcard               62.4  1238              5.95
```

```r
print(paste("Most effective treatment:", treatment_effects$treatment[1],
    "with", round(treatment_effects$effect_vs_control[1], 1),
    "percentage point increase"))
```

```
[1] "Most effective treatment: Personal Visit with 12.3 percentage point increase"
```

2.4 8. Cross-Sectional Survey Analysis

Dataset: political_attitudes_2024.csv

2.4.1 8.1 Cross-Sectional Exploration

```r
# Load the dataset
attitudes <- read_csv("political_attitudes_2024.csv")

# Examine the data
glimpse(attitudes)
```

```
Rows: 1,200
Columns: 11
$ respondent_id      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
$ age                <dbl> 36, 34, 32, 36, 44, 41, 81, 40, 60, 73, 69, 23, 87~
$ gender             <chr> "male", "female", "female", "female", "female", "f~
$ race_ethnicity     <chr> "White", "White", "Latino", "Other", "White", "Whi~
$ education          <dbl> 4, 5, 2, 2, 4, 1, 5, 3, 1, 3, 4, 2, 4, 1, 4, 4, 5,~
$ income_bracket     <dbl> 2, 10, 5, 10, 2, 1, 7, 4, 4, 7, 8, 7, 9, 10, 4, 6,~
$ ideology           <dbl> 5, 5, 3, 2, 6, 6, 5, 4, 4, 1, 3, 2, 4, 4, 6, 1, 3,~
```

```
$ party_id            <chr> "Republican", "Independent", "Independent", "Repub~
$ trust_gov           <dbl> 2, 2, 5, 0, 5, 6, 5, 4, 4, 4, 2, 4, 9, 3, 3, 4, 0,~
$ policy_support_env  <dbl> 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,~
$ policy_support_guns <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,~
```

```r
# Summary statistics
attitudes %>%
    summarise(mean_age = mean(age, na.rm = TRUE), mean_trust = mean(trust_gov,
        na.rm = TRUE), mean_env_support = mean(policy_support_env,
        na.rm = TRUE), mean_gun_support = mean(policy_support_guns,
        na.rm = TRUE))
```

```
# A tibble: 1 x 4
  mean_age mean_trust mean_env_support mean_gun_support
     <dbl>      <dbl>            <dbl>            <dbl>
1     53.2       4.04            0.717            0.321
```

2.4.2 8.2 Limitations of Cross-Sectional Data

```r
# What questions CAN this data answer?  1. Current
# associations between variables
correlation_trust_age <- cor(attitudes$age, attitudes$trust_gov,
    use = "complete.obs")
print(paste("Correlation between age and trust:", round(correlation_trust_age,
    3)))
```

```
[1] "Correlation between age and trust: -0.054"
```

```r
# 2. Group differences at one point in time
ideology_differences <- attitudes %>%
    group_by(ideology) %>%
    summarise(mean_env_support = mean(policy_support_env, na.rm = TRUE),
        mean_gun_support = mean(policy_support_guns, na.rm = TRUE),
        n = n())

print("Policy support by ideology:")
```

```
[1] "Policy support by ideology:"
```

```r
print(ideology_differences)
```

```
# A tibble: 7 x 4
  ideology mean_env_support mean_gun_support     n
     <dbl>            <dbl>            <dbl> <int>
1        1            0.921           0.0674    89
2        2            0.924           0.185    119
3        3            0.865           0.2      170
4        4            0.695           0.282    387
5        5            0.638           0.367    218
6        6            0.565           0.604    154
7        7            0.413           0.651     63
```

```r
# What questions CANNOT this data answer?
print("This cross-sectional data CANNOT answer:")
```

```
[1] "This cross-sectional data CANNOT answer:"
```

```r
print("1. How individuals' attitudes change over time")
```

```
[1] "1. How individuals' attitudes change over time"
```

```r
print("2. Whether age causes changes in trust (could be generational)")
```

```
[1] "2. Whether age causes changes in trust (could be generational)"
```

```r
print("3. Causal effects of education on ideology")
```

```
[1] "3. Causal effects of education on ideology"
```

```r
print("4. Whether policy attitudes are becoming more polarized")
```

```
[1] "4. Whether policy attitudes are becoming more polarized"
```

```r
# Example of correlation vs causation
education_trust <- attitudes %>%
    group_by(education) %>%
    summarise(mean_trust = mean(trust_gov, na.rm = TRUE), n = n())

print("Trust by education (correlation, not causation):")
```

```
[1] "Trust by education (correlation, not causation):"
```

```r
print(education_trust)
```

```
# A tibble: 5 x 3
  education mean_trust     n
      <dbl>      <dbl> <int>
1         1       4.00   250
2         2       4.14   225
3         3       3.97   246
4         4       4.00   237
5         5       4.08   242
```

```r
print("We cannot say education CAUSES trust differences - could be confounded by income, age
```

```
[1] "We cannot say education CAUSES trust differences - could be confounded by income, age, 
```