Week 3, Class 6: Practice Exercises

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 = ________, b = ________, c = ________, d = ________

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: ________

1.1.3 1.3 True or False: Research Designs

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

________ Experiments always require a laboratory setting ________ Natural experiments rely on events outside researcher control ________ Cross-sectional data can show individual change over time ________ Panel data follows the same people over multiple periods ________ 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 ________ creates quasi-random assignment
2. The researcher does not ________ the treatment
3. Groups become comparable by ________
4. We can compare ________ and control groups
5. The assignment process is ________ to political outcomes

Word bank: event, control, chance, treatment, unrelated

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: ________

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: ________

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 = ________, b = ________, c = ________, d = ________

1.3.2 3.2 Fill in the Code: Panel Data

Complete this code to analyze panel data:

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

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: ________

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: ________

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 = ________, b = ________, c = ________, d = ________

1.4.3 4.3 True or False: Research Tradeoffs

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

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

2 AI Exercises

Tips for Working with Claude:

- Ask for **R code using only tidyverse** (no other packages)
- Request **simple, focused answers** to your specific question—not complex analyses
- Ask Claude to **explain what the code is doing** since you're learning
- Avoid asking for visualizations or plots in these exercises
- Include the output of `glimpse()` in your prompt so Claude knows your variable names

Example prompt: "Using tidyverse in R, compare the mean poverty rate between counties where wage_change is TRUE vs FALSE. Keep the code simple and explain what each line does. Here is what my data looks like: [paste glimpse output]"

For each AI exercise: - Work with Claude to analyze the data - Record your prompts and key findings

2.1 5. Analyzing a Natural Experiment

Dataset: border_policy_change.csv

Description: Data from counties where minimum wage laws differ.

Variables: - `county_fips`: County FIPS code (int) - `year`: Year of observation (int) - `unemployment_rate`: Unemployment percentage (dbl) - `median_income`: Median household income (dbl) - `gini_index`: Income inequality measure (dbl) - `poverty_rate`: Poverty percentage (dbl) - `pop_density`: Population density (dbl) - `urban_rural`: Urban or rural classification (chr) - `percent_white`: Percentage white population (dbl) - `percent_black`: Percentage black population (dbl) - `percent_hispanic`: Percentage hispanic population (dbl) - `wage_change`: Whether minimum wage changed in that year (lgl)

2.1.1 5.1 Understanding the Natural Experiment

```r
# Load the dataset
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.5.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.2
v ggplot2   4.0.1     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
border_data <- read_csv("border_policy_change.csv")
```

```
Rows: 500 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): urban_rural
dbl (10): county_fips, year, unemployment_rate, median_income, gini_index, p...
lgl  (1): wage_change

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Examine the structure
glimpse(border_data)
```

```
Rows: 500
Columns: 12
$ county_fips       <dbl> 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4~
$ year              <dbl> 2018, 2019, 2020, 2021, 2022, 2018, 2019, 2020, 2021~
$ unemployment_rate <dbl> 7.365339, 11.196894, 3.519065, 9.348586, 5.821456, 7~
$ median_income     <dbl> 66804.75, 59666.68, 68788.24, 33449.04, 49137.52, 70~
$ gini_index        <dbl> 0.4983657, 0.4011999, 0.3540930, 0.3895261, 0.447380~
$ poverty_rate      <dbl> 15.582417, 13.355892, 8.278077, 12.625750, 10.510356~
$ pop_density       <dbl> 78.30621, 2323.19213, 217.33144, 33.25231, 10.00000,~
$ urban_rural       <chr> "Suburban", "Urban", "Suburban", "Urban", "Rural", "~
$ percent_white     <dbl> 59.04994, 82.20877, 75.29557, 69.55309, 38.29192, 53~
$ percent_black     <dbl> 16.095203, 37.405300, 38.099078, 34.732050, 37.14794~
$ percent_hispanic  <dbl> 27.150750, 38.602322, 5.916968, 40.622173, 19.632986~
$ wage_change       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALS~
```

2.1.2 5.2 Implementing the Design

Ask Claude to help you: - Compute poverty rates and unemployment in treated and control counties

2.2 6. Working with Panel Data

Dataset: voter_panel_study.csv

Description: Panel survey following the same voters over multiple elections.

Variables: - `respondent_id`: Unique voter identifier (int) - `wave`: Panel wave number (int) - `age`: Age at time of survey (int) - `education`: Education level (chr) - `ideology`: Political ideology (chr) - `political_interest`: Interest in politics (int) - `social_media_use`: Social media usage (chr) - `participated_protest`: Whether participated in protest (lgl) - `protest_issue`: Issue of protest if participated (chr) - `voted_last_election`: Whether voted in last election (lgl)

2.2.1 6.1 Exploring Panel Structure

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

```
Rows: 900 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (4): education, ideology, social_media_use, protest_issue
dbl (4): respondent_id, wave, age, political_interest
lgl (2): participated_protest, voted_last_election

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
Rows: 900
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> 18, 19, 20, 21, 22, 23, 26, 27, 28, 69, 70, 71, 3~
$ education            <chr> "Some College", "Some College", "Some College", "~
$ ideology             <chr> "Liberal", "Conservative", "Very Conservative", "~
$ political_interest   <dbl> 9, 1, 7, 4, 2, 7, 9, 8, 8, 9, 2, 8, 4, 3, 10, 10,~
$ social_media_use     <chr> "Never", "Always", "Sometimes", "Rarely", "Never"~
$ participated_protest <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, ~
$ protest_issue        <chr> NA, "Immigration", NA, NA, "Economy", NA, NA, NA,~
$ voted_last_election  <lgl> TRUE, FALSE, TRUE, FALSE, FALSE, TRUE, TRUE, TRUE~
```

2.2.2 6.2 Analyzing Change Over Time

Work with Claude to: - Track how individual respondents' ideology and political interest change across waves - Is average Analyze social media higher or lower for those who voted in the last election? - Identify voters who become more or less politically engaged over time based on voting behavior

2.3 7. Experimental Design Analysis

Dataset: gotv_experiment.csv

Description: Get-out-the-vote field experiment data.

Variables: - `voter_id`: Voter identifier (int) - `treatment`: Treatment assignment (chr) - `age_group`: Age group category (chr) - `education`: Education level (chr) - `voted_2022`: Whether voted in 2022 (lgl)

2.3.1 7.1 Understanding the Experiment

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

```
Rows: 1000 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): treatment, age_group, education
dbl (1): voter_id
lgl (1): voted_2022

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
Rows: 1,000
Columns: 5
$ voter_id   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ treatment  <chr> "Control", "Mail", "Canvass", "Phone", "Control", "Phone", ~
$ age_group  <chr> "30-44", "65+", "30-44", "30-44", "65+", "18-29", "18-29", ~
```

```
$ education  <chr> "Graduate", "High School", "Some College", "BA", "Graduate"~
$ voted_2022 <lgl> FALSE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, ~
```

2.3.2 7.2 Analyzing Treatment Effects

Ask Claude to help you: - Calculate turnout rates by treatment group - Compare the mean turnout rate for each group to control. Which had the greatest effect on turnout?

2.4 8. Cross-Sectional Survey Analysis

Dataset: political_attitudes_2024.csv

Description: Large cross-sectional survey of political attitudes.

Variables: - `respondent_id`: Unique identifier (int) - `age`: Age in years (int) - `gender`: Gender identity (chr) - `race_ethnicity`: Race/ethnicity (chr) - `education`: Highest degree (chr) - `income_bracket`: Income bracket (chr) - `ideology`: Political ideology (chr) - `party_id`: Party identification (chr) - `trust_gov`: Trust in government (int) - `policy_support_env`: Environmental policy support (int) - `policy_support_guns`: Gun policy support (int)

2.4.1 8.1 Cross-Sectional Exploration

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

```
Rows: 500 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (6): gender, race_ethnicity, education, income_bracket, ideology, party_id
dbl (5): respondent_id, age, trust_gov, policy_support_env, policy_support_guns

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Examine the data
glimpse(attitudes)
```

```
Rows: 500
Columns: 11
$ respondent_id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
$ age                 <dbl> 69, 52, 50, 44, 25, 20, 27, 58, 42, 22, 63, 66, 22~
$ gender              <chr> "Female", "Female", "Male", "Female", "Male", "Fem~
$ race_ethnicity      <chr> "Hispanic", "Hispanic", "Black", "Hispanic", "Blac~
$ education           <chr> "Graduate", "Some College", "Graduate", "Less than~
$ income_bracket      <chr> "<$25k", "<$25k", "<$25k", "$25-50k", "$75-100k", ~
$ ideology            <chr> "Conservative", "Moderate", "Conservative", "Very ~
$ party_id            <chr> "Independent", "Republican", "Democrat", "Independ~
$ trust_gov           <dbl> 7, 4, 3, 5, 10, 0, 9, 7, 4, 3, 0, 5, 9, 7, 3, 1, 9~
$ policy_support_env  <dbl> 4, 9, 9, 10, 2, 2, 4, 3, 9, 8, 6, 10, 4, 10, 7, 4,~
$ policy_support_guns <dbl> 10, 6, 1, 6, 3, 1, 8, 9, 9, 0, 5, 9, 6, 0, 2, 7, 1~
```

2.4.2 8.2 Limitations of Cross-Sectional Data

Work with Claude to: - Identify what questions this data can answer - Discuss what questions it cannot answer - Explore correlations vs causal claims