Problem Set 2: When Data Lies

Measurement, Distributions, and Scientific Fraud

Gov 51

Invalid Date

0.1 Instructions

In 2014, a graduate student named Michael LaCour published a blockbuster study in *Science* with senior co-author Donald Green. The study claimed that a single 20-minute conversation with a gay canvasser could durably change voters' attitudes toward same-sex marriage. The paper received enormous media attention and influenced political campaigns.

In 2015, two other graduate students — David Broockman and Joshua Kalla — discovered that the data were fabricated. Using the measurement tools you've been learning in this course, they showed that LaCour's data were statistically indistinguishable from a pre-existing national survey. *Science* retracted the paper.

In this problem set, you will retrace their steps. You'll start by analyzing LaCour's data as if it were real, then use histograms, summary statistics, and correlations to catch the fraud yourself, examine what real data looks like, and finally use regression to formalize your findings.

Expected time: Approximately 2.5–3 hours. Parts can be completed in separate sessions.

Submission Requirements:

1. Create a GitHub repository named `gov51-ps2`
2. Include folders: `data/raw/`, `code/`, `output/`
3. Place data files in `data/raw/`
4. Your rendered PDF should be in the root directory
5. Include the URL to your GitHub repository at the top of your submission
6. Make meaningful commits as you work (at least 5 commits)

Data files (download from course website):

- `gay.csv` — LaCour's experimental data on attitudes toward same-sex marriage (69,592 observations)
- `gayreshaped.csv` — The same data reshaped with feeling thermometer scores across survey waves
- `ccap2012.csv` — The 2012 Cooperative Campaign Analysis Project, a large national survey

1 Part 1: The Original Study (20 points)

In this section, you'll analyze LaCour's data as if the study were real. This is what researchers did when the paper first came out.

1.1 Exploring the Experiment

1.1.1 Load packages and data

```
# Your code here
library(tidyverse)

gay <- read_csv("data/raw/gay.csv")
```

Q1.1 (3 points): How many total observations are in `gay.csv`? How many unique values does `study` take? List all the treatment conditions in the `treatment` variable.

```
# Your code here
```

Q1.2 (3 points): Filter to Study 1, Wave 1 only. Calculate the mean of `ssm` (same-sex marriage attitude, scored 1–5) for each treatment group using `group_by()` and `summarize()`. Which treatment group shows the highest average support?

```
# Your code here
```

Q1.3 (3 points): Still using Study 1, Wave 1, calculate the standard deviation of `ssm` within each treatment group. Are the standard deviations roughly similar across groups? Why is this important in a randomized experiment? (2–3 sentences)

```
# Your code here
```

Your interpretation:

Q1.4 (3 points): Create a bar plot of the mean `ssm` by treatment group (Study 1, Wave 1) using `ggplot()` with `geom_col()`. *Hint:* First create a summary data frame with the means, then plot it.

```
# Your code here
```

1.2 Checking Balance and Comparing Outcomes

Q1.5 (4 points): A well-run randomized experiment should produce groups that look similar at baseline. Filter `gay.csv` to Study 1, Wave 1. Compare the "Same-Sex Marriage Script by Gay Canvasser" group and the "No Contact" group:

- Calculate the mean and SD of `ssm` for each group
- Compute the **standardized difference in means**: $\frac{\bar{x}_{\text{gay canvasser}} - \bar{x}_{\text{no contact}}}{\sqrt{(s^2_{\text{gay canvasser}} + s^2_{\text{no contact}})/2}}$

A standardized difference below 0.25 in absolute value suggests good balance. Do these groups look similar at baseline? (2–3 sentences)

```
# Your code here
```

Your interpretation:

Q1.6 (4 points): Now look at the outcomes. Filter to Study 1, **Wave 2** only. Calculate the mean `ssm` for the gay canvasser group and the No Contact group. What is the difference in means (gay canvasser minus no contact)?

Interpret: since the groups were balanced at baseline, what does a difference at Wave 2 suggest? (2–3 sentences)

```
# Your code here
```

Your interpretation:

2 Part 2: The Forensic Discovery (30 points)

Broockman, Kalla, and Aronow showed that LaCour fabricated his data — and they caught him using exactly the tools from this course: histograms, means, standard deviations, and correlations. You'll now retrace their steps.

2.1 The Baseline Distribution

Q2.1 (3 points): Load `gayreshaped.csv` and `ccap2012.csv`. Report the number of observations and column names for each dataset. In `gayreshaped.csv`, the variables `therm1` through `therm4` represent feeling thermometer scores (0–100) toward gay men and lesbians, measured at each survey wave.

```
# Your code here
gay_r <- read_csv("data/raw/gayreshaped.csv")
ccap <- read_csv("data/raw/ccap2012.csv")
```

Q2.2 (5 points): Extract the baseline feeling thermometer (`therm1`) for Study 1 from `gayreshaped.csv`. Calculate its mean, median, and standard deviation. Then do the same for `gaytherm` from `ccap2012.csv` (remove NAs). Report both sets of statistics side by side. What do you notice? (1–2 sentences)

```
# Your code here
```

What do you notice?

Q2.3 (5 points): Create side-by-side histograms comparing the LaCour Study 1 baseline (`therm1`) and the CCAP data (`gaytherm`). Use the same bin breaks for both: `breaks = seq(0, 100, by = 5)`.

Hint: Combine the two datasets into one data frame with a column indicating the source, then use `ggplot()` with `facet_wrap()`.

Do these look like they came from different populations? (2–3 sentences)

```
# Your code here
```

Your interpretation:

2.2 Re-test Reliability

When real people take the same survey twice, their answers correlate — but not perfectly. People's moods change, they round differently, they misread questions. A re-test correlation of 0.95–0.97 is normal for feeling thermometers. A correlation near 1.0 is suspicious.

Q2.4 (6 points): For both the Study 1 and Study 2 **control groups** ("No Contact"), calculate:

(a) The correlation between `therm1` and `therm2` (use `cor()` with `use = "complete.obs"`)
(b) The standard deviation of the within-person change (`therm2 - therm1`)

Present your results side by side. A re-test correlation of 0.95–0.97 is typical for real feeling thermometer data. Which study shows a more plausible pattern, and why? (3–4 sentences)

```
# Your code here
```

Your interpretation:

Q2.5 (5 points): Create two scatter plots side by side:

(a) Study 1 control group: `therm1` (x-axis) vs `therm2` (y-axis), with a 45-degree line
(b) Study 2 control group: `therm1` (x-axis) vs `therm2` (y-axis), with a 45-degree line

Describe the differences. Which looks more like real longitudinal data, and why? (3–4 sentences)

```
# Your code here
```

Your interpretation:

2.3 Putting It All Together

Q2.6 (6 points): You've now found three irregularities:

1. The baseline distributions match an existing national survey
2. The re-test correlations are implausibly high (Study 1)
3. The within-person changes are implausibly small (Study 1)

In 4–5 sentences, explain why these three findings together suggest the data were fabricated rather than collected in a real experiment. Use specific numbers from your analysis. Why would a fabricator produce data with these properties?

Your interpretation:

3 Part 3: What Real Data Looks Like (25 points)

You've found irregularities in Study 1. Now analyze Study 2 and see if the same patterns hold.

3.1 Comparing Baselines

Q3.1 (3 points): Filter `gayreshaped.csv` to Study 2. How many observations? Calculate the mean and SD of `therm1` (baseline) for the treatment group and the control group separately.

```
# Your code here
```

Q3.2 (3 points): Create side-by-side histograms of Study 1 baseline (`therm1`) and Study 2 baseline (`therm1`), using the same bins as before. How do they compare visually? (2–3 sentences)

```
# Your code here
```

Your interpretation:

3.2 Checking Balance and Comparing Outcomes in Study 2

Q3.3 (3 points): Filter `gayreshaped.csv` to Study 2, keeping only the "Same-Sex Marriage Script by Gay Canvasser" and "No Contact" groups. Check balance at baseline: calculate the mean and SD of `therm1` for each group and the standardized difference in means (same formula as Q1.5). Do the groups look balanced? (1–2 sentences)

```
# Your code here
```

Your interpretation:

Q3.4 (3 points): Now compare outcomes. Calculate the mean `therm2` (post-conversation) for each group. What is the difference in post-period means (gay canvasser minus no contact)?

```
# Your code here
```

3.3 Uncertainty Around the Difference

Q3.5 (4 points): Calculate the standard error of the difference in post-period means from Q3.4. Use the formula:

$$SE = \sqrt{\frac{s^2_{\text{gay canvasser}}}{n_{\text{gay canvasser}}} + \frac{s^2_{\text{no contact}}}{n_{\text{no contact}}}}$$

where s^2 is the variance of `therm2` and n is the sample size for each group. Then construct a 95% confidence interval: $(\bar{x}_1 - \bar{x}_2) \pm 1.96 \times SE$. Does the CI include zero? What does this tell you about whether the groups differ after the conversations?

```
# Your code here
```

Your interpretation:

3.4 Tracking Attitudes Over Time

Q3.6 (4 points): For the Study 2 **control group**, calculate the correlation between `therm1` and each subsequent wave: `therm2`, `therm3`, and `therm4`. Use `cor()` with `use = "complete.obs"`.

Do the correlations decay over time (i.e., do more distant waves correlate less with baseline)? What does this pattern tell you about the stability of real survey measurements? (2–3 sentences)

```
# Your code here
```

Your interpretation:

Q3.7 (5 points): You've now analyzed fabricated data (Study 1), caught the fraud, and examined more realistic data (Study 2). In 5–6 sentences, discuss:

- What makes Study 2's data different from Study 1's data? Be specific about which statistics differ and by how much.
- Why is it important that real treatment effects are smaller and messier than fabricated ones?
- What role did simple measurement tools (mean, SD, histograms, correlations) play in detecting the fraud?

Your reflection:

4 Part 4: Formalizing the Evidence with Regression (25 points)

Now formalize your findings using regression.

4.1 Regression as a Fraud Detector

Q4.1 (5 points): For the Study 1 **control group**, fit a linear regression predicting `therm2` from `therm1`:

$$\text{therm2}_i = \beta_0 + \beta_1 \times \text{therm1}_i + \epsilon_i$$

Store the result as `fit_s1` and display the summary. Report the slope ($\hat{\beta}_1$), the R-squared, and the residual standard error.

Hint: Use `lm(therm2 ~ therm1, data = ...)` and `summary()`.

```
# Your code here
```

Q4.2 (5 points): Repeat Q4.1 for the Study 2 **control group**. Store as `fit_s2`. Report the same three quantities: slope, R-squared, and residual standard error.

Compare the two models. In fabricated data, we'd expect $\hat{\beta}_1 \approx 1$ and $R^2 \approx 1$ because the "follow-up" data is just a copy of the baseline with tiny noise added. In real data, we'd expect regression to the mean: $\hat{\beta}_1 < 1$ and lower R^2. Which pattern do you see? (3–4 sentences)

```
# Your code here
```

Your interpretation:

4.2 Comparing Groups with Regression

Q4.3 (5 points): Now use regression to compare the two groups in Study 2. Create a binary variable `treated` that equals 1 for the "Same-Sex Marriage Script by Gay Canvasser" group and 0 for the "No Contact" group. Then fit:

$$\text{therm2}_i = \beta_0 + \beta_1 \times \text{treated}_i + \epsilon_i$$

Report $\hat{\beta}_1$, its standard error, t-statistic, and p-value. Compare $\hat{\beta}_1$ to the difference in means you calculated in Q3.4 — what do you notice? What does $\hat{\beta}_0$ estimate? Is the difference between groups statistically significant at $\alpha = 0.05$?

Note: Filter to only the gay canvasser and no contact groups before running the regression.

```
# Your code here
```

Your interpretation:

4.3 Controlling for Baseline Attitudes

Q4.4 (5 points): Add `therm1` as a control variable:

$$\text{therm2}_i = \beta_0 + \beta_1 \times \text{treated}_i + \beta_2 \times \text{therm1}_i + \epsilon_i$$

Report the coefficient on `treated`, its standard error, t-statistic, and p-value. Compare this to your results from Q4.3:

(a) Did the estimate of $\hat{\beta}_1$ change much?
(b) Did the standard error change? In which direction?
(c) Why does controlling for baseline attitudes improve precision? (2–3 sentences)

```
# Your code here
```

Your interpretation:

Q4.5 (5 points): Compare the R-squared values from Q4.3 and Q4.4. Why does adding `therm1` increase the R-squared so dramatically? What does this tell you about how much of the variation in follow-up attitudes is explained by where people started versus the treatment?

Finally, use `predict()` to generate predicted values from your Q4.4 model for two hypothetical respondents:

- Person A: `treated = 1`, `therm1 = 50` (moderate baseline)
- Person B: `treated = 0`, `therm1 = 50` (same baseline, no treatment)

What is the predicted difference in their Wave 2 attitudes?

```
# Your code here
```

Your interpretation:

5 Summary

In this problem set, you:

1. **Used measurement tools** — means, standard deviations, histograms, and correlations — to detect fabricated data
2. **Used regression** to formalize group comparisons and showed that regression with a binary predictor gives the same answer as the difference in means
3. **Distinguished real from fabricated data** using the tools from this course

The LaCour case shows that you don't need advanced statistics to catch fraud. The tools you've been learning all semester were enough for two graduate students to overturn a *Science* paper.

Submission Checklist:

☐ GitHub repository URL at the top
☐ All code chunks run without error
☐ All interpretation questions answered in complete sentences
☐ At least 5 meaningful commits
☐ PDF renders cleanly

Data sources:

- LaCour, Michael J., and Donald P. Green. 2014. "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality." *Science* 346(6215): 1366–1369. (Retracted)
- Broockman, David, Joshua Kalla, and Peter Aronow. 2015. "Irregularities in LaCour (2014)." Working paper.
- Cooperative Campaign Analysis Project (CCAP). 2012.