Detecting Fraud and Testing Hypotheses

Gov 51 Section — Week 5

George

Harvard University

February 25, 2026

Today's Plan

Part 1: LaCour Forensics (~35 min)

- Load real survey data (ANES 2016)
- Simulate fabrication step by step
- QQ plots and KS tests
- See how the fraud was detected

Part 2: Hypothesis Testing (~35 min)

- Same ANES data, new question
- Sample mean, SE, t-statistic
- p-value and confidence intervals
- `t.test()` in R

Today is **hands-on R throughout**. Open RStudio and follow along.

Part 1: The LaCour Forensics

The American National Election Study

The **ANES** is one of the longest-running academic surveys in the US, conducted every election year since 1948.

One signature question is the **feeling thermometer**:

> "How would you rate *[group]*? Ratings between 50 and 100 mean you feel favorable toward them. Ratings between 0 and 50 mean you feel unfavorable. 50 means you feel neither favorable nor unfavorable."

Today we'll use the 2016 ANES thermometer toward **gay men and lesbians** — the same kind of data LaCour claimed to have collected.

Load the ANES thermometer data

```
library(tidyverse)
anes <- read_csv("anes_thermometer.csv")
nrow(anes)
## [1] 3598
```

3,598 respondents each gave a rating from 0 to 100.

This is **real** survey data. We'll use it as our baseline to understand what fabricated data looks like by comparison.

Step 1: Start with real data as your baseline

```r
set.seed(51)

# LaCour's method: sample from real survey data
fake_control <- sample(anes$thermometer,
                       size = 1000,
                       replace = TRUE)

# Compare means
mean(anes$thermometer)   ## [1] 60.73
mean(fake_control)       ## [1] ~60.7
```

If you just resample from real data, the distributions are nearly identical. That's step 1 of the fabrication recipe.

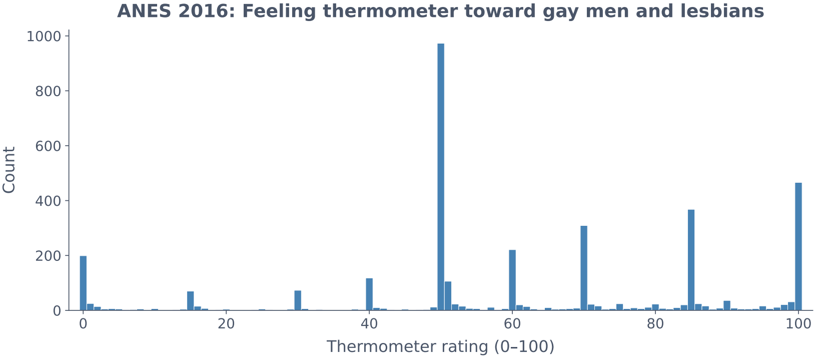
What does real thermometer data look like?

The ANES asked: “Rate gay men and lesbians on a scale from 0 to 100.” Let’s see what people actually said:

```r
ggplot(anes, aes(x = thermometer)) +
  geom_histogram(binwidth = 1, fill = "steelblue") +
  labs(x = "Thermometer (0-100)", y = "Count",
       title = "ANES 2016: Gay men and lesbians")
```

Run this code now. What do you notice about the shape?

This is what real survey data looks like



People round to multiples of 5 and 10 — about 27% said exactly 50. Big spikes at 0, 50, and 100. This "heaping" is the **fingerprint of real human respondents**.

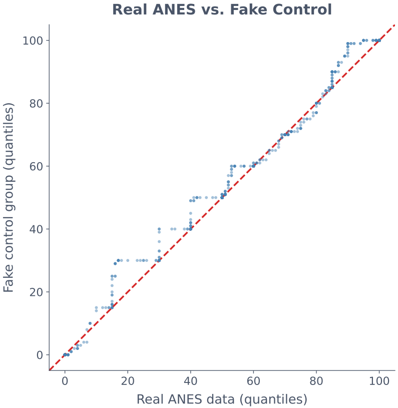
QQ plot reveals suspicious similarity

```r
qqplot(anes$thermometer, fake_control,
       main = "Real ANES vs. Fake Control",
       xlab = "Real ANES data",
       ylab = "Fake control group")
abline(0, 1, col = "red", lwd = 2)
```

- Points fall almost perfectly on the 45-degree line
- Real survey data from two **different** studies would never match this well

Broockman & Kalla's key insight: the distributions matched **too** perfectly.

The points fall right on the 45-degree line



Real data from two **different** studies would never match this well. This tipped off Broockman & Kalla.

The KS test confirms these distributions are suspiciously identical

```
ks.test(anes$thermometer, fake_control)
##
## Two-sample Kolmogorov-Smirnov test
## D = ~0.02, p-value = ~0.9
```

- **KS test**: maximum difference between empirical CDFs
- $D \approx 0$ means the distributions are nearly identical
- High p-value: cannot reject that they come from the same distribution

This is **suspicious**, not reassuring. Independent surveys should differ somewhat.

Turn to your neighbor

If you wanted to fabricate data that looked like a real survey ***but with a treatment effect****, what would you do to these numbers?*

Take 2 minutes. We'll discuss as a class.

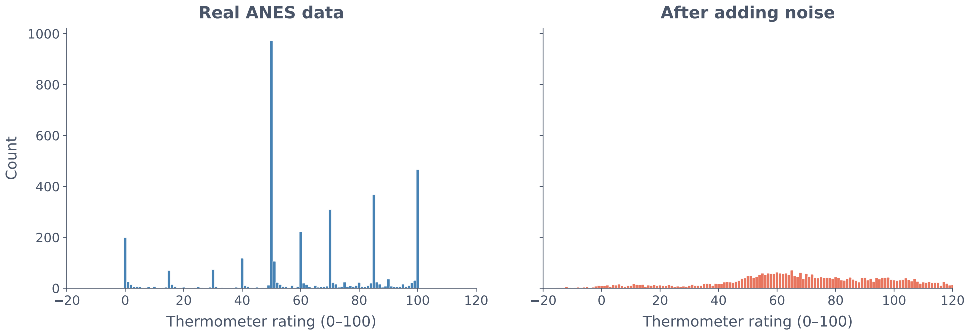
Step 2: Add normal noise to simulate a treatment effect

```r
noise <- rnorm(nrow(anes), mean = 8, sd = 10)
fabricated <- anes$thermometer + noise
```

```r
par(mfrow = c(1, 2))
hist(anes$thermometer, breaks = 50,
     main = "Real ANES", col = "steelblue",
     xlim = c(-20, 120))
hist(fabricated, breaks = 50,
     main = "After adding noise", col = "coral",
     xlim = c(-20, 120))
```

Run the code above — you should see the heaping disappear.

Adding noise smooths out the heaping



- ▷ The heaping at multiples of 5 is **gone**
- ▷ But: some values are now < 0 or > 100

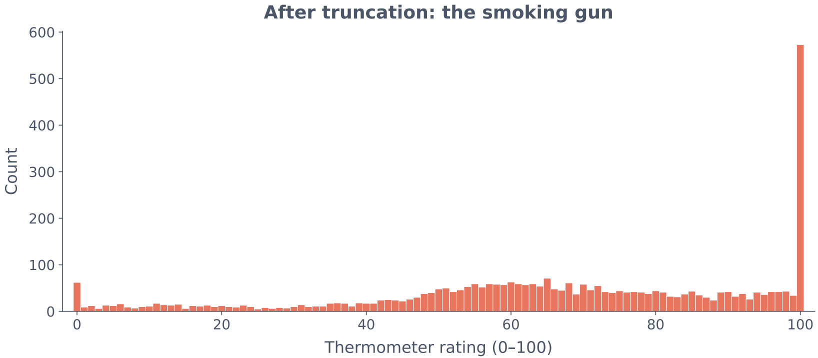
Step 3: Truncate — and create new artifacts

```r
fabricated_trunc <- pmin(pmax(fabricated, 0), 100)
sum(fabricated_trunc == 0)    # Pile-up at 0
sum(fabricated_trunc == 100)  # Pile-up at 100
```

```r
hist(fabricated_trunc, breaks = 50,
     main = "Truncated: the smoking gun",
     col = "coral")
```

Run it — look at what happens at the boundaries.

Truncation creates a brand new artifact



Pile-ups at 0 and 100 that weren't in the original. The natural heaping at 50 is gone. This is the smoking gun.

The forensic fingerprint of fabrication

1. **Real data** has heaping at 0, 50, 100, and multiples of 5
2. **Adding noise** smooths out natural heaping, makes the distribution approximately normal
3. **Truncating** creates artificial spikes at boundaries (0 and 100) that weren't there before

LaCour's data had exactly this signature: no heaping at 50, excess mass at boundaries, and distributions that matched the CCAP baseline too perfectly.

Data fabrication leaves fingerprints. Real surveys have predictable artifacts. Forensic statistics can detect fraud.

Part 2: Hypothesis Testing

Is the average rating different from neutral?

You already loaded the data. Now a new question:

Research question: Is the average feeling thermometer rating toward gay men and lesbians different from 50 (neutral)?

```r
x <- anes$thermometer
n <- length(x)
cat("n =", n)
## n = 3598
```

50 is the midpoint of the 0–100 scale. If Americans are truly neutral on average, the mean should be near 50.

Step 1: Calculate the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```r
x_bar <- mean(x)

cat("x-bar =", round(x_bar, 2))
## x-bar = 60.73
```

The sample average is 60.73 — about 10.7 points above neutral.

Turn to your neighbor: How would you interpret 60.73 in plain English?

Step 2: How precise is our estimate?

$$\text{SE} = \frac{s}{\sqrt{n}}$$

```
s <- sd(x)
se <- s / sqrt(n)
cat("s =", round(s, 2))          ## s = 27.36
cat("SE =", round(se, 2))        ## SE = 0.46
```

- $s = 27.36$: individual ratings vary enormously (0 to 100)
- $\text{SE} = 0.46$: the **mean** is pinned down very precisely
- Why so small? Because $n = 3{,}598$ — SE shrinks with $\sqrt{n}$

Step 3: How many SEs away from 50?

$$t = \frac{\bar{x} - \mu_0}{\text{SE}} = \frac{\bar{x} - \mu_0}{s \,/\, \sqrt{n}}$$

```r
mu_0 <- 50
t_stat <- (x_bar - mu_0) / se
cat("t =", round(t_stat, 2))  ## t = 23.53
```

Our sample mean is **23.5 standard errors** above 50. That’s enormous.

Rule of thumb: $|t| > 2$ is “statistically significant.” Here $t = 23.5$ — not even close to ambiguous.

Turn to your neighbor

Based on $t = 23.5$:

1. Do you think we will reject $H_0 : \mu = 50$?
2. Will the 95% CI contain 50?
3. In lecture, the commute data gave $t = 1.14$ and *failed* to reject. Why is this result so different?

Take 2 minutes. Explain your reasoning to your partner.

Step 4: What's the probability of data this extreme?

$$p\text{-value} = 2 \times P(T > |t|), \quad T \sim t_{n-1} = t_{3597}$$

```r
p_val <- 2 * pt(-abs(t_stat), df = n - 1)
cat("p-value =", p_val)
## p-value = ~0 (R prints: < 2.2e-16)
```

$p \approx 0$ — we **reject** H_0 overwhelmingly. The data provide extremely strong evidence that the average feeling toward gay men and lesbians is *not* neutral.

Compare to lecture: commute data had $p = 0.26$ (fail to reject). Same method, different data, different conclusion.

Step 5: Where does the true mean plausibly lie?

$$\text{CI} \quad = \quad \bar{x} \quad \pm \quad t^*_{0.975,\, df} \times \text{SE}$$

```
t_crit <- qt(0.975, df = n - 1)
ci_lower <- x_bar - t_crit * se
ci_upper <- x_bar + t_crit * se
cat("95% CI:", round(ci_lower,2), "to", round(ci_upper,2))
## 95% CI: 59.84 to 61.63
```

- 50 is **far outside** $[59.84,\ 61.63]$ — consistent with rejecting
- CI is narrow (< 2 points) because $n = 3{,}598$
- CI and test always agree: μ_0 outside CI $\Rightarrow$ reject

Turn to your neighbor

A friend says: "The 95% CI is $[59.84, 61.63]$*,*
so there's a 95% chance the true mean is in that interval."

Is your friend right? Why or why not?

Take 2 minutes. This is one of the most commonly misunderstood ideas in statistics.

R does it all in one line

```
t.test(x, mu = 50)
##      One Sample t-test
## t = 23.531, df = 3597, p-value < 2.2e-16
## 95 percent confidence interval:
##  59.84  61.63
## sample estimates:
## mean of x
##     60.73
```

Everything matches our by-hand calculations:
$t = 23.5, \quad p < 0.001, \quad \text{CI} = [59.84, \quad 61.63].$

Understand the pieces first, then use the shortcut.

Hypothesis testing: the five steps

Step	What	Formula	R code
1	Mean	$\bar{x} = \frac{1}{n} \sum x_i$	`mean(x)`
2	SE	$\text{SE} = s/\sqrt{n}$	`sd(x) / sqrt(n)`
3	t-stat	$t = (\bar{x} - \mu_0)/\text{SE}$	`(mean(x) - 50) / se`
4	p-value	$2 \times P(T > \lvert t \rvert)$	`2*pt(-abs(t), df=n-1)`
5	95% CI	$\bar{x} \pm t^*_{0.975} \times \text{SE}$	`t.test(x, mu = 50)`

Hypothesis testing asks: could this result be due to chance? The t-statistic measures signal relative to noise. The CI and p-value always tell the same story.

Wrapping Up

Both halves used the same data — and the same logic

- **Part 1 (LaCour):** Distributions, histograms, QQ plots — tools from weeks 2–3 used to detect fraud
- **Part 2 (ANES):** Mean, SE, *t*-test, CI — the core of hypothesis testing from this week's lecture
- Both parts: **Always look at your data.** Summary statistics alone can mislead.

PS2 is due **Thursday, March 5**. Start early!

Three things to do before next section

1. Finish Problem Set 2 (due Thursday, March 5)
2. Review: What does a p-value mean? What does it *not* mean?
3. Practice: Can you run `t.test()` on any numeric variable?

Questions?