Lab 4: Analysis of U.S. High School Education Rates

Ben Atkinson

March 2025

Introduction

This lab investigates changes in high school education rates across U.S. states between 2005 and 2018. The dataset includes education rates for each state in both years and indicates which region (South, West, Northeast, or Midwest) each state belongs to. Our key questions are:

(a) Did the high school education rate in the US improve between 2005 and 2018?

(b) Did the high school education rate in the South improve over this same time span?

(c) Are the 2018 education rates in the Northeast and the Midwest different or the same?

1 National Comparison: 2005 vs. 2018

Visualization

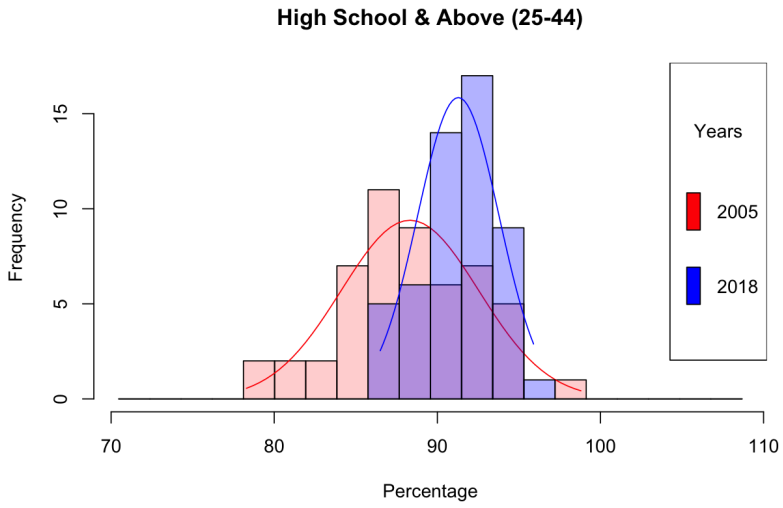

Figure 1: Histogram of High School & Above (25–44) for 2005 (red) and 2018 (blue).

Paired T-Test Results

We used a paired t-test to compare 2005 vs. 2018 rates:

```
paired_sample_a <- t.test(data_a1, data_a2, paired = TRUE)
```

The results were:

- **Test statistic (t):** -9.772999
- **Degrees of freedom (df):** 51
- **p-value:** 1.576538e-10
- **95% CI:** [-3.71, -2.22]
- **Mean difference:** -2.965385

Since the p-value is extremely small (1.58×10^{-10}), we reject the null hypothesis of no difference. The negative mean difference indicates that 2018 has a *higher* high school education rate than 2005 by roughly 2.97 percentage points.

2 Regional Analysis: The South

Normality Checks

A Shapiro-Wilk test was used to check normality:

- **2005 (South)**: Shapiro-Wilk W = 0.974, p = 0.909 (fail to reject normality)
- **2018 (South)**: Shapiro-Wilk W = 0.877, p = 0.034 (reject normality)

Because the 2018 data was non-normal, we used a Wilcoxon signed-rank test.

QQ Plots (Non-Normal vs. Normal)

.

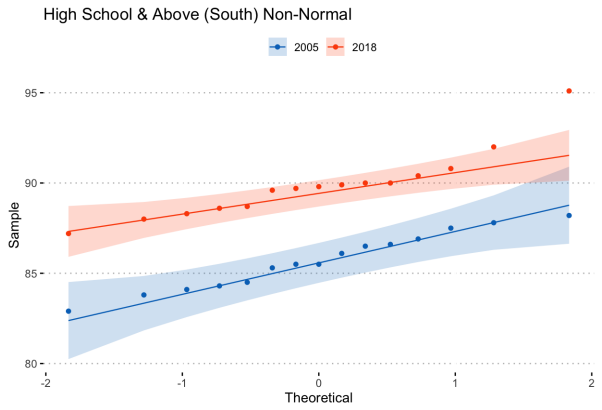

Figure 2: Non-Normal Q-Q Plot for South Data (2005 vs. 2018)

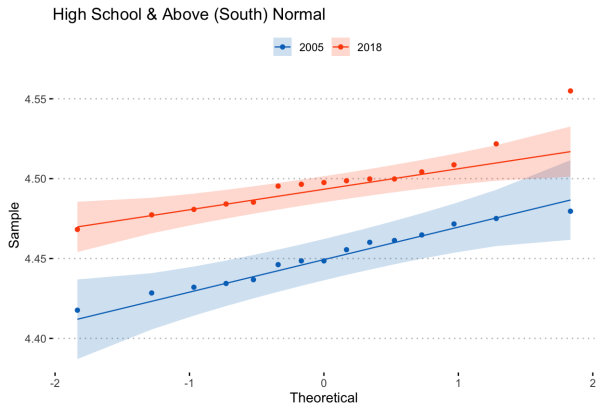

Figure 3: Log-Transformed Q-Q Plot for South Data (2005 vs. 2018)

Wilcoxon Signed-Rank Test

```
wilcox_test_d <- wilcox.test(data_d1, data_d2, paired = TRUE)
```

- **Test statistic (V):** 0
- **p-value:** 0.000721

The Wilcoxon test indicates a statistically significant difference between 2005 and 2018 rates in the South.

Log Transformation and Paired T-Test

We also took a log transformation of the South data and applied a paired t-test to it:

```
paired_sample_d <- t.test(data_log_d1, data_log_d2, paired = TRUE)
```

The log-transformed paired t-test results:

- **Test statistic (t):** -11.4
- **Degrees of freedom (df):** 7
- **p-value:** 1.79e-08
- **95% CI:** [-0.0564, -0.0386]
- **Mean difference:** -0.0475

The negative mean difference on the log scale again indicates that the 2018 rate is higher than the 2005 rate in the South, confirming the Wilcoxon results.

3 Regional Comparison: Northeast vs. Midwest in 2018

Visualization

We compared 2018 education rates in the Northeast (NE) vs. the Midwest (MW):

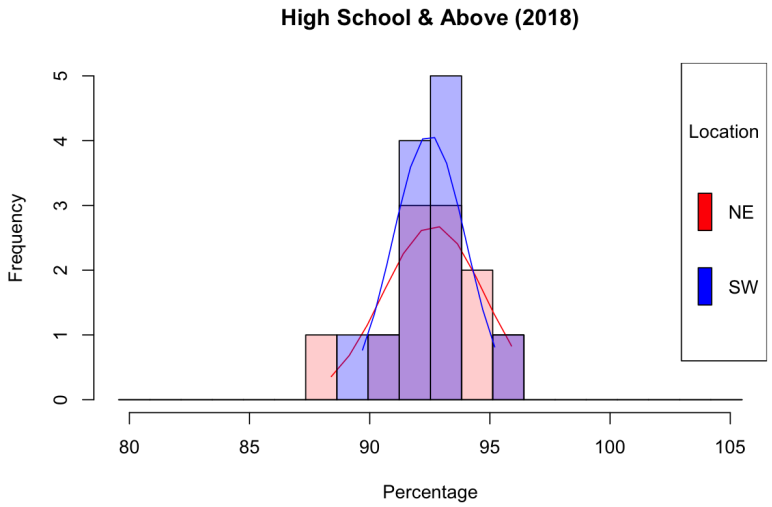

Figure 4: High School & Above (2018) for NE (red) vs. SW (blue).

Two-Sample T-Test

A two-sample t-test (assuming equal variance):

```
two_sample_c <- t.test(data_c1, data_c2, var.equal = TRUE, paired = FALSE)
```

The results:

- **Test statistic (t):** 0.235
- **Degrees of freedom (df):** 21
- **p-value:** 0.816
- **95% CI:** [-1.41, 1.77]
- **Mean difference:** -0.13

Since $p = 0.816$, we fail to reject the null hypothesis of no difference. The Northeast and Midwest 2018 rates appear to be statistically the same.

Conclusion

Based on the above tests, we conclude:

(a) **National Improvement**: The U.S. high school education rate showed a significant increase from 2005 to 2018 (paired t-test, $p \approx 1.58 \times 10^{-10}$).

(b) **Improvement in the South**: Both non-parametric (Wilcoxon) and log-transformed paired t-tests confirm that the South significantly improved its high school education rate from 2005 to 2018.

(c) **Northeast vs. Midwest**: The two-sample t-test for 2018 indicates no statistically significant difference in education rates between the Northeast and the Midwest ($p = 0.816$).