Does Test Prep Improve Math Test Scores

Ali Efe Isik

Introduction

Many high school students take test prep courses because they want to improve their academic performance, but whether test prep is associated with meaningful improvement is not always clear. In this report, I use the Kaggle "Students' Performance in Exams" dataset ($n = 1000$) to study math achievement and test-prep participation in a broader population of similar students. This question is relevant to students, families, and educators because test prep can take substantial time and money, and schools may use evidence like this when deciding whether to recommend or support prep programs. The main question is: "Do students who complete a test preparation course tend to score higher in math than students who do not?" To express this comparison, let X be the math score of a randomly selected student who completed test prep and let Y be the math score of a randomly selected student with no test prep, and define the difference random variable $D = X - Y$. I estimate the typical value of D using two estimators, the mean of differences and the median of differences, and then use resampling methods to compare these estimators and answer the research question.

Data description

The dataset used in this report contains 1,000 observations, where each row represents a unique high school student. It includes student demographic information, whether the student completed a test preparation course, and exam scores in math, reading, and writing. The population of interest is students similar to those represented in this Kaggle dataset who take standardized exams. The single random variable analyzed in this report is Math Score, a numerical value bounded between 0 and 100.

```
dat <- read.csv("~/Documents/Math-438/StudentsPerformance.csv", stringsAsFactors = TRUE)
```

```
# Print head
head(dat)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

```
# Pull out Math Score values for each test-prep group
# "completed" = students who completed a test preparation course
# "none"      = students who did not take a test preparation course
prep_math   <- dat$math.score[dat$test.preparation.course == "completed"]
```

```r
noprep_math <- dat$math.score[dat$test.preparation.course == "none"]

#remove NA's
prep_math   <- prep_math[!is.na(prep_math)]
noprep_math <- noprep_math[!is.na(noprep_math)]
```

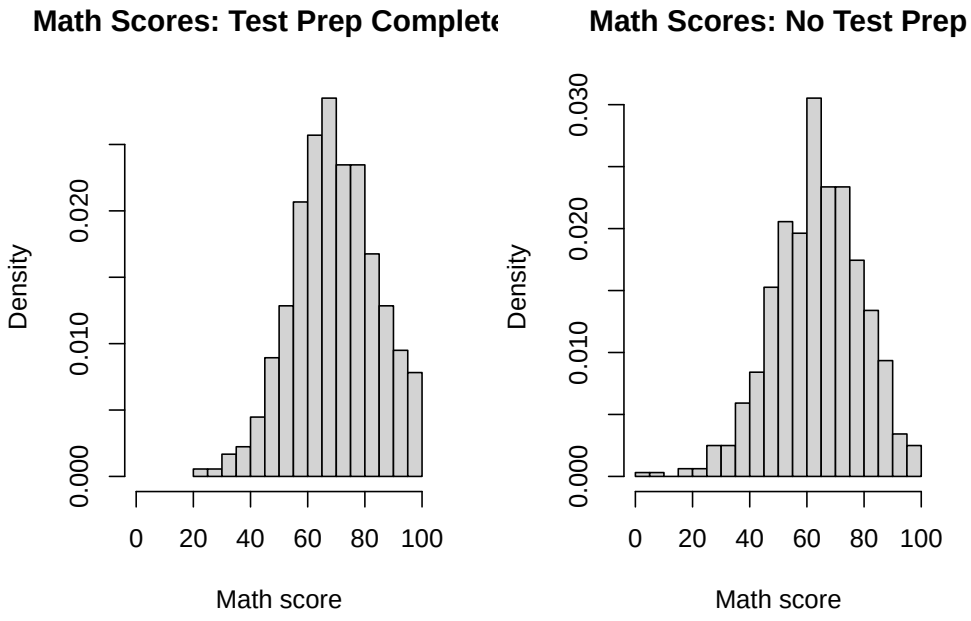
Preliminary distributions

After separating the prep and non-prep groups, I looked at the distributions of math scores in each group, which helps show why a simple Normal model may not be appropriate.

```r
# Plot math score distributions for each group on the same 0-100 scale
par(mfrow = c(1,2))

hist(prep_math, breaks = 20, freq = FALSE,
     xlim = c(0, 100),
     main = "Math Scores: Test Prep Completed",
     xlab = "Math score")

hist(noprep_math, breaks = 20, freq = FALSE,
     xlim = c(0, 100),
     main = "Math Scores: No Test Prep",
     xlab = "Math score")
```




```r
par(mfrow = c(1,1))
```

Both groups have math scores between 0 and 100, and neither histogram looks perfectly symmetric like a bell curve. A Normal distribution might seem like a reasonable model at first because test scores are numeric and often cluster around a middle value, but the Normal model is not appropriate here because the scores are bounded (0–100) and the histograms show departures from a bell shape. In addition, the full dataset is not one i.i.d. sample from a single distribution because it combines two different groups (completed vs. none), which can have different score distributions. In this analysis, I treat students within each group as

independent observations, but this assumption may be violated if students are clustered within the same schools or classrooms. Because of these features, I avoid relying on a Normal distribution model and instead use bootstrap resampling within each group to approximate the sampling distributions of my estimators.

Estimators

I compare two estimators of the central tendency of D. The first estimator is the mean of differences, $T_1 = \text{mean}(D)$, which estimates the average score advantage. The second estimator is the median of differences, $T_2 = \text{median}(D)$, which estimates the typical score advantage and is more robust to skewness and outliers. Both estimators target the same general population characteristic—the typical size of the score difference between the completed and no-prep groups—but summarize it in slightly different ways.

Because the prep and no-prep groups are separate samples with no natural one-to-one matching, I form random cross-group pairs to create a sample of differences that approximates draws of $D = X - Y$. Using these paired differences, I compute the observed mean of differences and observed median of differences as baseline estimates before applying the bootstrap procedure.

```
# Use the smaller group size so we can form paired differences of equal length
n_pair <- min(length(prep_math), length(noprep_math))

# Create one "observed" set of paired differences by randomly pairing students
# across the two groups (without replacement) for a baseline estimate.
set.seed(438)
x_obs <- sample(prep_math,   size = n_pair, replace = FALSE)  # completed group
y_obs <- sample(noprep_math, size = n_pair, replace = FALSE)  # no-prep group
D_obs <- x_obs - y_obs

# Observed estimators (based on the observed paired differences)
T1_obs <- mean(D_obs)      # mean of differences
T2_obs <- median(D_obs)    # median of differences

# Print labeled results
cat("Observed mean of differences (T1_obs)   =", round(T1_obs, 3), "\n")
```

```
## Observed mean of differences (T1_obs)   = 5.704
```

```
cat("Observed median of differences (T2_obs) =", round(T2_obs, 3), "\n")
```

```
## Observed median of differences (T2_obs) = 6.5
```

Bootstrap method

To see how much these estimators could change from sample to sample, I use a nonparametric bootstrap. In each bootstrap run, I resample math scores with replacement from the completed group and resample the same number of scores with replacement from the no-prep group. I subtract the resampled scores to create a bootstrap set of differences $D^* = X^* - Y^*$, then compute the mean of differences $T_1^* = \text{mean}(D^*)$ and the median of differences $T_2^* = \text{median}(D^*)$. Repeating this many times gives empirical sampling distributions for T_1 and T_2, which lets me compare their variability and uncertainty.

$D^* = X^* - Y^*, \qquad T_1^* = \text{mean}(D^*), \qquad T_2^* = \text{median}(D^*)$

```
# Bootstrap (sampling distributions of the estimators)
set.seed(438)
B <- 1000

T1_boot <- numeric(B)
T2_boot <- numeric(B)
```

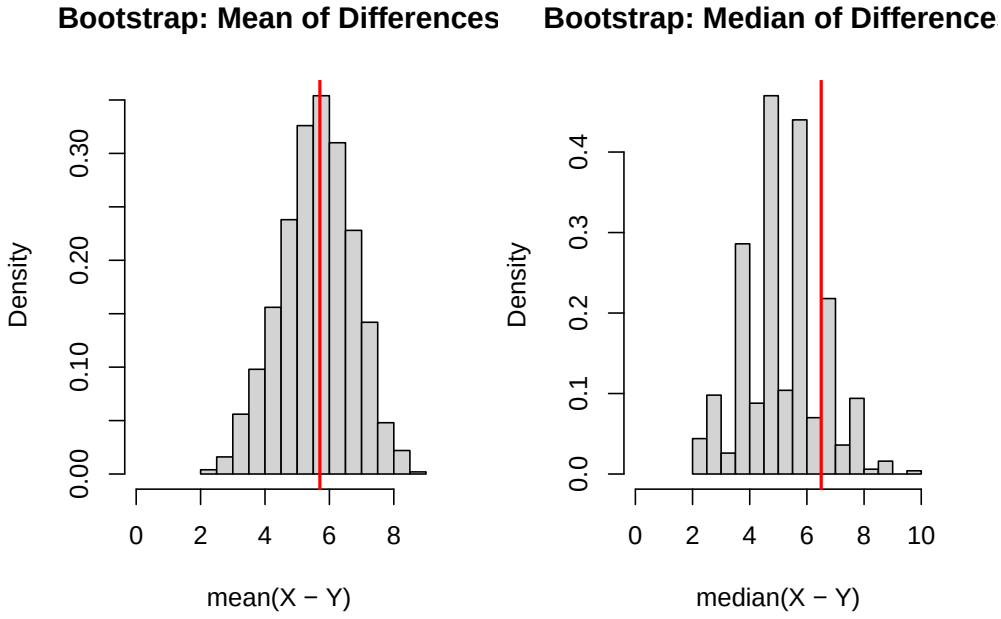
```r
for (b in 1:B) {
  # Resample within each group (with replacement)
  x_star <- sample(prep_math,   size = n_pair, replace = TRUE)
  y_star <- sample(noprep_math, size = n_pair, replace = TRUE)

  # Form paired differences and compute estimators
  D_star <- x_star - y_star
  T1_boot[b] <- mean(D_star)
  T2_boot[b] <- median(D_star)
}
```

```r
# Plot bootstrap sampling distributions of the two estimators
par(mfrow = c(1,2))

# Mean of differences: include 0 and ensure we don't cut off the right tail
hist(T1_boot, breaks = 20, freq = FALSE,
     xlim = c(0, max(T1_boot, T1_obs)), #ChatGPT
     main = "Bootstrap: Mean of Differences",
     xlab = "mean(X - Y)")
abline(v = T1_obs, col = "red", lwd = 2)

# Median of differences: include 0 and ensure we don't cut off the right tail
hist(T2_boot, breaks = 20, freq = FALSE,
     xlim = c(0, max(T2_boot, T2_obs)), #ChatGPT
     main = "Bootstrap: Median of Differences",
     xlab = "median(X - Y)")
abline(v = T2_obs, col = "red", lwd = 2)
```




```r
par(mfrow = c(1,1))
```

Results

The bootstrap histograms show the empirical sampling distributions of the mean of differences and the median of differences. Each histogram is centered near the corresponding observed estimate, shown by the red vertical line, and the spread of each distribution represents how much the estimator would vary across repeated samples. Visually, the more concentrated histogram corresponds to the estimator that is more stable.

```
# Bootstrap means and standard errors (SE = sd of bootstrap values)
mean_T1_boot <- mean(T1_boot)
se_T1_boot   <- sd(T1_boot)

mean_T2_boot <- mean(T2_boot)
se_T2_boot   <- sd(T2_boot)

# 95% percentile confidence intervals
CI_T1 <- quantile(T1_boot, probs = c(0.025, 0.975))
CI_T2 <- quantile(T2_boot, probs = c(0.025, 0.975))

# Create a summary table (ChatGpt)
results_table <- data.frame(
  Statistic   = c("T1: mean of differences", "T2: median of differences"),
  Observed    = c(T1_obs, T2_obs),
  Boot_Mean   = c(mean_T1_boot, mean_T2_boot),
  Boot_SE     = c(se_T1_boot,   se_T2_boot),
  CI_Lower_95 = c(CI_T1[1], CI_T2[1]),
  CI_Upper_95 = c(CI_T1[2], CI_T2[2])
)

# Rounded for display (ChatGpt)
results_table_rounded <- results_table
results_table_rounded[ , -1] <- round(results_table_rounded[ , -1], 3)

results_table_rounded
```

```
##                   Statistic Observed Boot_Mean Boot_SE CI_Lower_95 CI_Upper_95
## 1   T1: mean of differences    5.704     5.612   1.114       3.363       7.752
## 2 T2: median of differences    6.500     5.410   1.375       3.000       8.000
```

To compare the estimators, I use the bootstrap standard errors and 95% confidence intervals shown in the table above. The mean of differences estimator (T_1) has a smaller bootstrap standard error (1.114 vs. 1.375) and a narrower 95% interval [3.363, 7.752] than the median of differences estimator (T_2), which has a 95% interval [3.000, 8.000]. This indicates that T_1 is more stable across repeated samples and provides a more precise estimate of the typical score advantage. For that reason, I use T_1 to answer the research question.

Conclusion

Using the mean of differences as the preferred estimator, the estimated typical math-score advantage for students who completed test prep is about 5.6 points. In other words, when comparing a randomly selected student who completed test prep to a randomly selected student who did not, the completed student is expected to score roughly 5–6 points higher in math on average. The 95% bootstrap percentile confidence interval for this advantage is [3.363, 7.752]. Because the entire interval is above 0, the results provide evidence that the population-level average difference in math scores between the completed and no-prep groups is positive. Overall, within the population represented by this dataset, completing a test preparation course is associated with higher math performance by several points on a 0–100 scale.

Limitations

This dataset is observational, so the results may not generalize broadly beyond students similar to those in this sample. Students who complete test prep may differ from those who do not in other ways (motivation, prior achievement, resources, etc.), which could explain part of the difference.

References

Adeyemi, Timothy. "Students Performance in Exams." Kaggle Datasets. https://www.kaggle.com/datasets/timothyadeyemi/students-performance-in-exams

Acknowledgements

I used ChatGPT to help debug RMarkdown knitting issues and adjust R chunks, and Grammarly for grammar/wording editing. All statistical choices, computations, and interpretation were done by me.