STAT 2600 - Final Project

Dr. Kris Pruitt

16 January 2025

Overview

Prior to 2011, the Colorado Department of Education (CDE) employed the Colorado Student Assessment Program (CSAP) to evaluate student performance in reading, writing, and math. State-level standardized testing evolved in subsequent years, but the CSAP data remains available via the CDE Information Marketplace at this link. The site also includes a plethora of metadata regarding the source material and the variables included in each table.

For this project, CDE administrators are interested in reviewing the academic performance of 9th graders (high school freshmen) between 2009 and 2010. During this time, the test results for mathematics were particularly low across the State of Colorado. However, increased focus on math curriculum for beginning high school students did improve student performance from 2009 to 2010. Consequently, administrators hope to better understand the transition in test scores between years and apply any valuable insights to future initiatives.

Suppose you have been contracted by the CDE to perform an in-depth analysis of the 2009-2010 CSAP results for 9th graders. Specifically, administrators seek answers to the following questions:

- Was there truly an improvement in passing rates for the math exam between 2009 and 2010? If so, was there a decline in performance on the reading and writing exams?
- Was there an association between a school's passing rate on the math exam and its passing rates on the reading and writing exams? If so, how accurately could math passing rates have been predicted from the reading and writing rates?

Your goal is to write a professional-quality report to answer these two questions by applying the 5A Method described throughout the textbook. Using the case studies in Chapters 3-5 as exemplars, your report must describe each of the five steps of the problem solving process with a particular focus on the exploratory, inferential, and predictive analyses required to answer the administrators' questions. When complete, your report must meet the following criteria:

- Generated in RMarkdown and submitted as a pdf file
- No more than 10 pages in length
- Begin with title, author, and date
- Include section headers for the 5A Method steps
- Label figures, tables, and page numbers
- Does not "echo" the R code, but does display results
- Written in complete sentences and paragraphs
- Peer-reviewed for spelling, grammar, and clarity

Your submission should be a stand-alone document that does not require additional context on the part of the reader. Your target audience (stakeholder) comprises CDE administrators who have an introductory level of understanding in statistics. In the sections that follow, we outline the key tasks that must be completed at each step of the 5A Method.

Ask a Question

Set the stage for your project by offering a brief background on the problem and restating the administrators' questions, in your own words. Describe the value of answering the question and the audience that might be concerned.

Acquire the Data

Describe the process required to import and wrangle the project data. You do not need to provide administrators a step-by-step guide to your process. You just need to make sure they understand where the data was obtained, why it is reputable, and how you prepared it for analysis, in general. Below, we offer some guidance to get you started.

The CDE provides public access to schools' test results via the Information Marketplace. After navigating to the page, click on the blue "Action" button on the upper right and select "API". A pop-up then provides the API Endpoint necessary to access the data. In order to overcome the 1,000-observation limit imposed by the site, you will need to import the data in three phases. For each of the three imports, you will add filters to the API request to isolate the appropriate class year and a specific test type. Add the following lines to the end of the API Endpoint and generate the associated data frames using techniques demonstrated in Chapter 2.2 of the textbook.

- ?grade_level=9&subject=READING
- ?grade_level=9&subject=WRITING
- ?grade_level=9&subject=MATH

After import, you should have three tables each with 637 observations. Combine the three tables into a single data frame with 1,911 total rows. This aggregated data frame comprises the *population* of all 9th grade test scores in the State of Colorado for the 2009 and 2010 school years. Next, you will perform a variety of wrangling tasks using techniques demonstrated in Chapter 2.3 of the textbook. Complete all of the following steps in order.

- Eliminate any rows with a `school_no` equal to zero.
- Filter out any rows that do not have at least 31 total students in both 2009 (`_09_total_count`) *and* 2010 (`_10_total_count`).
- Limit the test result columns (7-30) to only those representing counts (i.e., `_count`).
- Assign the numeric data type to all variables except subject. Define subject as a factor.
- Create a numeric variable for year by moving its value from the columns (`_09` and `_10`) to rows (`2009` and `2010`). This will require renaming all of the count variables to eliminate the year indicator.
- Limit the columns to only year, school_no, subject, and the five test results (not scored, unsatisfactory, partially proficient, proficient, and advanced).

After wrangling, a `glimpse()` of the data frame should indicate the following structure.

```
## Rows: 1,488
## Columns: 8
## $ year           <dbl> 2009, 2009, 2009, 2010, 2010, 2010, 2009, 2009, 2009, 2~
## $ school_no      <dbl> 10, 10, 10, 10, 10, 10, 15, 15, 15, 15, 15, 15, 24, 24,~
## $ subject        <fct> MATH, READING, WRITING, MATH, READING, WRITING, MATH, R~
## $ noscore        <dbl> 0, 1, 0, 4, 2, 3, 0, 0, 0, 0, 0, 0, 18, 17, 16, 12, 9, ~
## $ unsatisfactory <dbl> 351, 141, 68, 313, 113, 83, 37, 8, 2, 32, 9, 9, 287, 76~
## $ partial        <dbl> 98, 195, 339, 139, 222, 348, 36, 20, 43, 48, 23, 56, 81~
## $ proficient     <dbl> 34, 151, 81, 51, 174, 78, 20, 66, 44, 31, 86, 50, 32, 1~
## $ advanced       <dbl> 5, 0, 0, 6, 1, 0, 4, 3, 8, 10, 3, 6, 4, 1, 2, 7, 1, 3, ~
```

Each observation now represents a unique combination of year, school number, and test subject. There are 248 different schools, each with six observations (2 years times 3 subjects). Consequently, your final data frame should consist of 1,488 rows and 8 columns.

As a final step prior to analysis, you will randomly sample 120 schools from the population using **your 9-digit student identification number** as the seed. Below is the code you will need to obtain your unique list of schools. The example code assumes that the full data frame is named `schools_all` and that the student ID is `123456789`. Replace the notional student ID with your own.

```
#create list of unique school numbers
school_list <- unique(schools_all$school_no)

#randomly select 120 schools
set.seed(123456789)
my_school_list <- sample(x=school_list,size=120,replace=FALSE)

#filter on selected schools
my_schools <- schools_all %>%
  filter(school_no %in% my_school_list)
```

For the remainder of the project, your data frame (`my_schools`) will be different than that of any other student. It should consist of 720 rows (6 for each school) and 8 columns. This data frame comprises your *sample* of all 9th grade test scores in the State of Colorado for the 2009 and 2010 school years.

Analyze the Data

Conduct inferential and predictive analyses to gather evidence related to the two research questions. At a minimum, your analytic results should include *five* plots (described below) and an explanation of their statistical significance.

To start, you will compare the proportions of passing (proficient and advanced) students between 2009 and 2010 for each test subject. This task requires reshaping the `my_schools` data frame and computing the difference in proportions between years. As an example, the difference in proportions for *writing* is shown below. This data is based on the notional randomization seed of `123456789` from the previous task.

```
## Rows: 120
## Columns: 4
## $ school_no <dbl> 15, 24, 40, 76, 146, 187, 203, 206, 210, 212, 263, 298, 370,~
## $ rate2009  <dbl> 0.5360825, 0.1658768, 0.2083333, 0.7442623, 0.9285714, 0.350~
## $ rate2010  <dbl> 0.46280992, 0.19393939, 0.13846154, 0.71052632, 0.83050847, ~
## $ diff      <dbl> -0.073272557, 0.028062617, -0.069871795, -0.033735979, -0.09~
```

For each of the 120 schools, we now have the difference in passing rate (`diff`) from 2009 to 2010. A negative value indicates a decrease in passing rate, while a positive value indicates an increase. The mean of the difference column is -0.034, suggesting the average school in Colorado witnessed a roughly 3 percentage-point decrease in passing rate for writing. However, this value is a point estimate based on a single sample of high schools. In order to better understand the variability of this statistic, we require a sampling distribution and a confidence interval.

Figure 1 depicts the bootstrap sampling distribution for the mean difference in *writing* pass rate from 2009 to 2010, along with 95% confidence bounds. This graphic is constructed using techniques demonstrated in Chapter 4.2 of the textbook. Given that zero is *not* within the confidence interval, it appears that the passing rate for writing got worse from 2009 to 2010. Specifically, we are 95% confident that the passing rate for writing decreased by between 2.0 and 4.9 percentage-points for 9th graders at the average Colorado school. For your report, you will need to repeat this analysis for all three subjects using your unique sample data. This will allow you to answer the administrators' first question.

Finally, you will model the association between passing rates in order to predict the performance in math based on the other two subjects. In particular, you will train a model on the 2009 passing rates and test its accuracy on the 2010 passing rates. After splitting the `my_schools` data frame by year, you must reshape

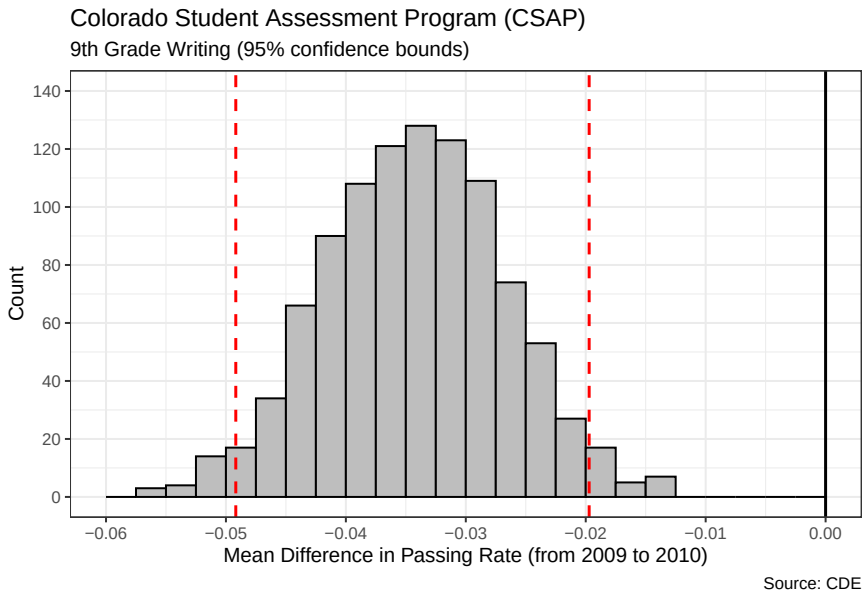

Figure 1: Mean difference in passing rate for writing from 2009 to 2010.

each table to move the subject passing rates from rows to columns. For example, a `glimpse()` of your new training set should appear as follows.

```
## Rows: 120
## Columns: 4
## $ school_no <dbl> 15, 24, 40, 76, 146, 187, 203, 206, 210, 212, 263, 298, 370,~
## $ MATH      <dbl> 0.24742268, 0.08530806, 0.05208333, 0.55081967, 0.55357143, ~
## $ READING   <dbl> 0.7113402, 0.3530806, 0.3541667, 0.8918033, 0.9642857, 0.505~
## $ WRITING   <dbl> 0.5360825, 0.1658768, 0.2083333, 0.7442623, 0.9285714, 0.350~
```

Each of the 120 schools has a single row with columns for the passing rate of each subject in 2009. After reshaping the training set, you must visualize the association between each pair of subjects. For example, Figure 2 displays the association between passing rates for reading and writing in the training set.

Based on this plot, the association between the passing rates for reading and writing is linear, increasing, and strong. This makes intuitive sense, because schools with students who perform better in writing would be expected to also perform better in reading. The correlation coefficient of 0.97 confirms the strength of the relationship.

The slope of the best-fit line (red) appears close to 1 (dashed black line), but slightly more shallow. Additionally, the intercept of the line is clearly greater than 0. By fitting a linear model to the training data and extracting the estimated parameters, we confirm the visual findings and obtain the following regression equation.

$$\text{READING} = 0.217 + 0.869 \cdot \text{WRITING}$$

The positive intercept suggests that students tend to perform better in reading than writing. However, a slope of less than one indicates this gap in performance gets smaller as the passing rate for writing gets larger. For example, a writing rate of 0.20 predicts a reading rate of 0.39 (difference of 0.19). By contrast, a writing rate of 0.80 predicts a reading rate of 0.91 (difference of 0.11). Thus, the difference in performance between the two subjects decreases as the writing performance increases.

In order to estimate the accuracy of this model, we use it to predict the 2010 data and compute the root mean squared error (RMSE). This method relies on techniques demonstrated in Chapter 5.2 of the textbook.

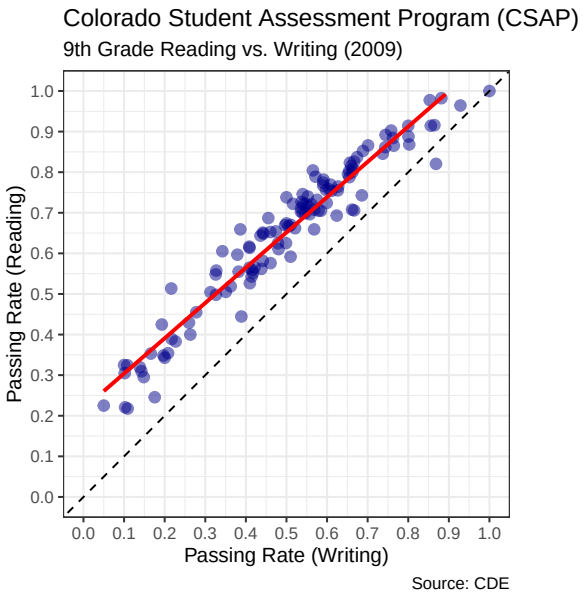

Figure 2: Association between passing rates for reading and writing.

```
## # A tibble: 1 x 1
##     RMSE
##    <dbl>
## 1 0.0677
```

Thus, if we know a school's passing rate for writing, then we can predict the passing rate for reading to within about 6.8 percentage points. For your report, you will need to repeat this analysis with math passing rates on the y-axis and reading/writing rates on the x-axis. This will allow you to answer the administrators' second question.

Advise on Results

Interpret the results of your quantitative analysis in qualitative terms. In other words, explain the practical significance of your analysis in the context of standardized testing and the evaluation of student performance. Employ language and terminology that is appropriate for the CDE administrators. Discuss the limitations of your results and how your insights could inform future initiatives.

Answer the Question

Directly answer the two questions posed by administrators. Recommend future analysis that could provide further insights or extend the applicability of your work.