PTT Analysis of Entrance Exam Scores in Taiwan

Christine P. Chai

May 23, 2020

I just started the template.

Executive Summary

Write something here

Introduction

I am curious about the relationship between the high school entrance exam score percentiles and the college entrance scores in Taiwan. It seems obvious that a greater percentage of students from top high schools get admitted to prestigious universities[1].

However, the high school entrance exam is much easier than the college entrance exam, so some people studied little in middle school and was able to get into a good high school. Then some of these people kept studying little and ended up with a bad score on the college entrance exam. On the other hand, I have also seen some students from mediocre high schools worked very hard during the three years, and eventually earned a stellar score in the college exam.

Therefore, I decided to gather data and create my own analysis. The target audience of this document would be students taking Statistics 101.

Background

In Taiwan, the high school entrance exam score percentiles are between 1% and 99%, and people often refer to the percentile rank (PR value) as from 1 to 99. This scoring system existed from 2001 to 2013[2]. The actual exam score ranges were different. For example, the maximum possible score was originally set to 300, but it was increased to 312 in Year 2007. Then the maximum possible score was increased to 412 in Year 2009. Therefore, the percentile ranks (PR values) serves as a normalized tool to compare academic achievements across different years.

The college entrance exams are held twice a year in Taiwan. The first exam, typically held in late January or early February, is called the General Scholastic Ability Test (GSAT)[3]. The second exam is called the Advanced Subjects Test (AST)[4], and it is almost always held on July 1st, 2nd, and 3rd. The GSAT scores are normalized to a range of 0 to 75, regardless of the difficulty level of GSAT each year. On the other hand, the scores of AST can vary widely because each subject is scored separately from 0 to 100. Since the AST scores fluctuate more due to the difficulty level of the exam questions each year, I decided to use the GSAT scores as a benchmark of the college exam scores.

Remark: The GSAT consists of five subjects, each of which are graded on a 0 to 15 point scale. Starting in 2019, students may choose four of the five subjects for the GSAT. The maximum possible score (i.e., full marks) is reduced from 75 to 60.

[1] https://bit.ly/2JSPXKc
[2] https://bit.ly/2JNQaOI
[3] https://bit.ly/2W0fdUq
[4] https://bit.ly/2J7YxoW

Data Description

It is a challenge to obtain individual pairs of data as a representative sample. Although it is easy to send out a spreadsheet and ask my friends to report their scores anonymously, this approach can result in a large selection bias. Many of my friends graduated from the same high school or college as I did, so we are likely to have similar entrance exam scores.

Hence I retrieved data from the SENIORHIGH (high school)[5] discussion section on PTT[6], the largest terminal-based bulletin board in Taiwan. I assume the data to be more representative (than if I had collected on my own) because anyone can get a PTT account and reply to the post. The majority of scores were reported in May 2015, and a few scores were reported in the following month or later.

The data *ptt_SENIORHIGH_data.csv* contain 197 rows, and the main variables are:

- **pttID**: Each person's ID on PTT, which can be anonymous. This column serves as the unique identifier of each person.
- **HighSchool_PR**: Each person's percentile rank (PR) of the high school entrance exam in Taiwan, ranging from 0 to 99.
- **College_Score**: Each person's General Scholastic Ability Test (GSAT) score, ranging from 0 to 75.

There are 6 missing values in **HighSchool_PR** and 3 missing values in **College_Score**, so I recorded each of them as "-1" (an invalid value).

In some cases, the reported scores can be inaccurate based on the respondent's description, so I created two indicators for this issue:

- **HS_Inacc**: A "1" means the reported score of high school entrance exam is inaccurate.
- **College_Inacc**: A "1" means the reported score of college entrance exam is inaccurate.

For instance, some people reported their percentile rank (PR) from the mock exam, rather than the actual high school entrance exam. Since there are two college entrance exams in each school year, some people may do much better on the second exam than the first one. Then they were admitted to a more prestigious school than the first exam score had indicated, so this is also a form of inaccuracy.

Raw data

Showing the first 10 rows of data.

```
data = read.csv("ptt_SENIORHIGH_data.csv")
names(data)[1] = "pttID"

data[1:10,]
```

```
##           pttID HighSchool_PR College_Score HS_Inacc College_Inacc
## 1    game275415            60            50        1            NA
## 2      a2654133            60            52       NA            NA
## 3   cookie20125            99            72       NA            NA
## 4       heejung            92            54        1            NA
## 5         shun01            87            51       NA            NA
## 6     robinyu85            -1            74       NA            NA
## 7    allengoose            69            48       NA            NA
## 8  godpatrick11            98            60       NA            NA
## 9     morgankhs            95            65       NA            NA
## 10      jazzard            88            65       NA            NA
```

[5] https://www.ptt.cc/bbs/SENIORHIGH/M.1432729401.A.995.html

[6] If you have a PTT account, you can log into the website using a browser. https://iamchucky.github.io/PttChrome/?site=ptt.cc

Exploratory Data Analysis

The first step in a data project is exploratory data analysis, before we perform any statistical modeling. Therefore, I start with observing the trends of the two main variables, **HighSchool_PR** and **College_Score**.

High School Entrance Exam Scores (Percentile Rank)

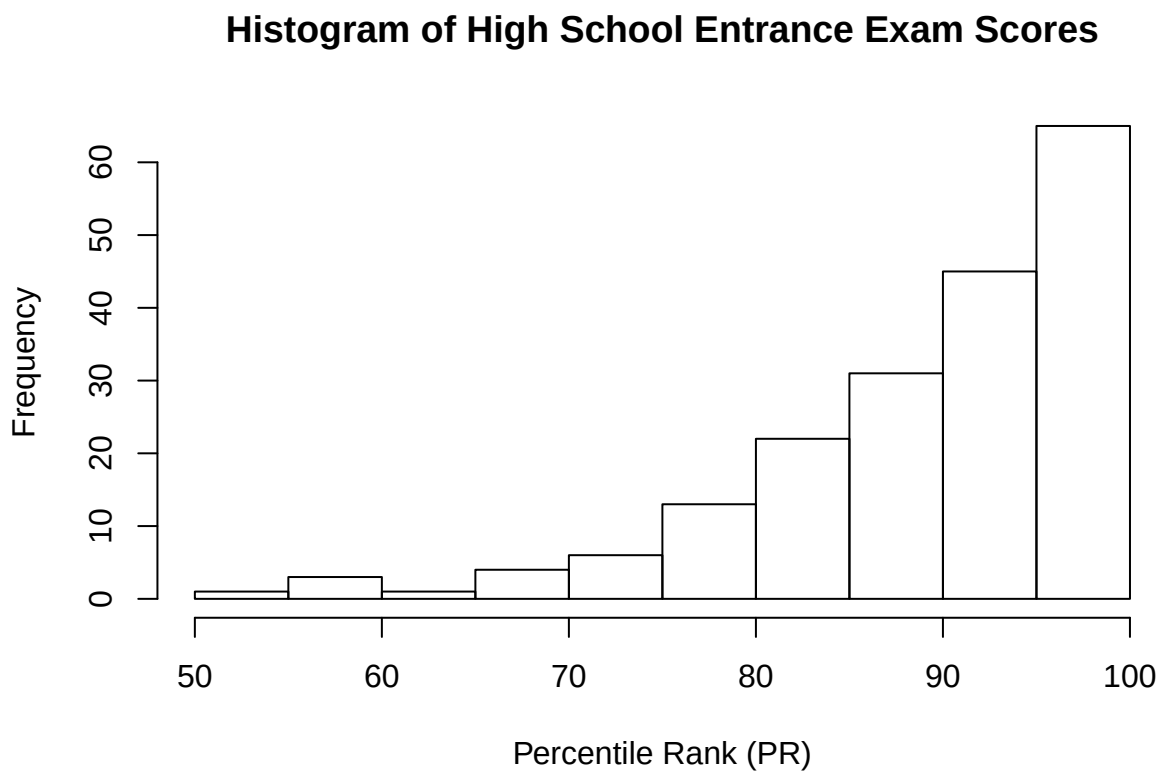
Below shows the descriptive statistics of **HighSchool_PR**, i.e., the percentile rank of high school entrance exam scores. The missing values are removed beforehand. Approximately 75% of the respondents have a percentile rank (PR) at least 85, indicating that most of the respondents scored in the top 15% of the high school entrance exam. The histogram is also extremely left-skewed.

```
# High school entrance exam scores: Remove missing values
uni_HS_score = data$HighSchool_PR[which(data$HighSchool_PR != -1)]

summary(uni_HS_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.00   85.00   92.00   89.82   97.00   99.00
```

```
hist(uni_HS_score, main = "Histogram of High School Entrance Exam Scores",
     xlab="Percentile Rank (PR)")
```



College Entrance Exam Scores

Similarly, I also show the descriptive statistics of **College_Score**, i.e., the college entrance exam scores between 0 and 75. The histogram is also left-skewed, but less extreme than **HighSchool_PR**.

According to the reference score table[7] from Wikipedia, the 88th percentile of the college entrance score fluctuates around 60 in Years 2004-2010, and 62-65 in Years 2011-2018. Since the median of **College_Score** is 64.5, we can infer that at least 50% of the respondents scored in the top 12% of the college entrance exam.

On the other hand, the reference score table also shows that the 75th percentile of the college entrance score is between 53 and 58 in Years 2004-2018. The PTT data's 1st quantile is already at 58, so we can also infer that at least 75% of the respondents scored in the top 25% of the college entrance exam.

Unfinished below

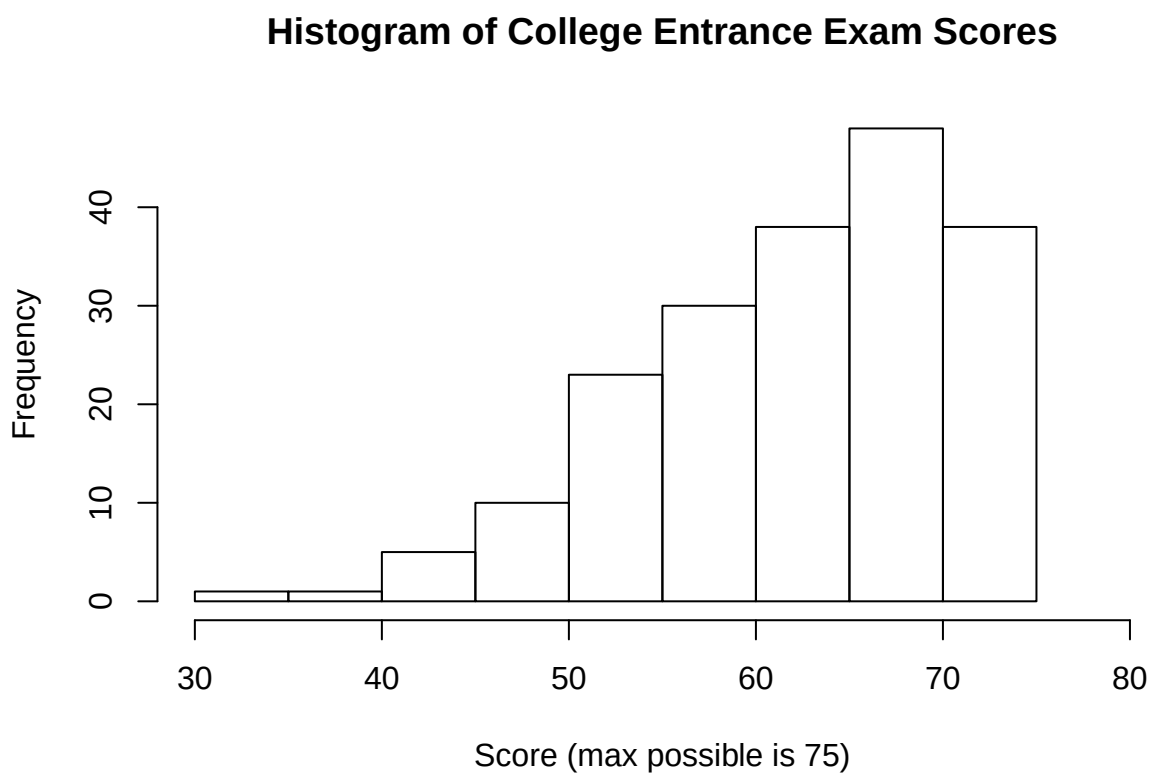
Write an inference paragraph: What are the general demographics of students who use PTT?

```
# College entrance exam scores: Remove missing values
uni_college_score = data$College_Score[which(data$College_Score != -1)]

summary(uni_college_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    34.0    58.0    64.5    62.7    69.0    75.0
```

```
hist(uni_college_score, main = "Histogram of College Entrance Exam Scores",
     xlab="Score (max possible is 75)",xlim=c(30,80))
```



Bivariate

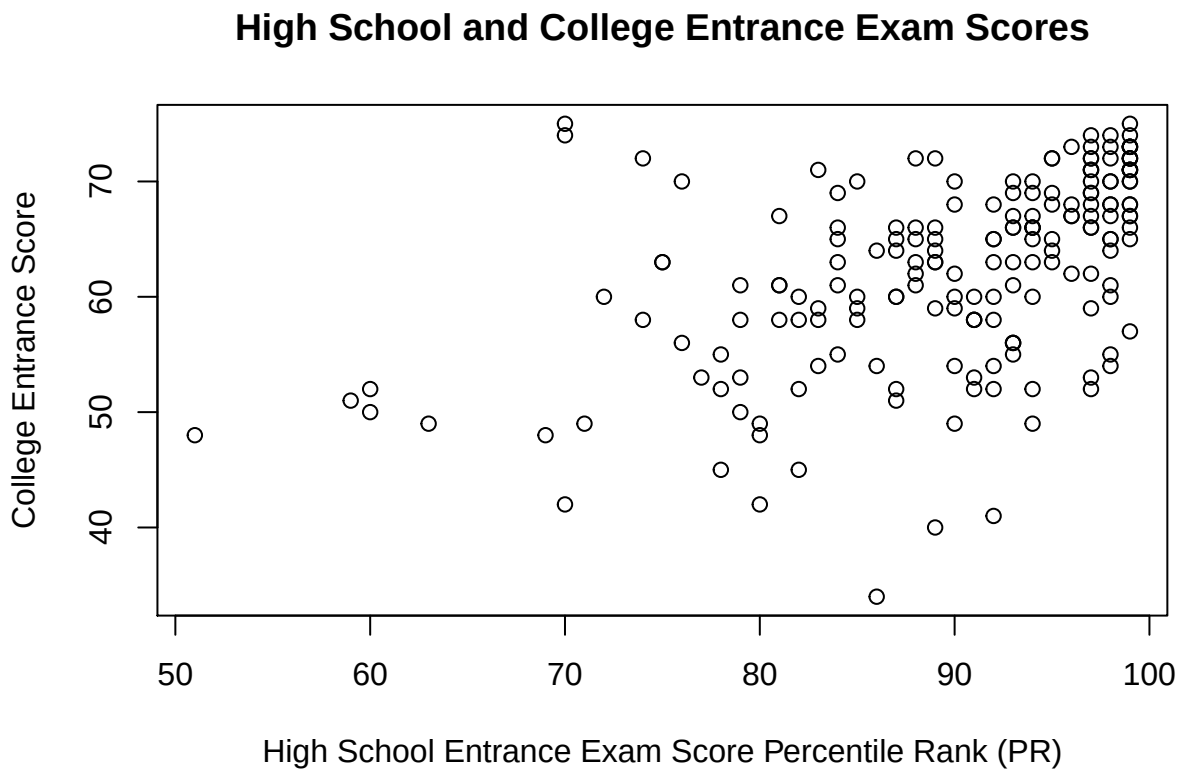
Then bivariate plot.

[7] https://bit.ly/3bAYOvO

```r
missing_rows = which(data$HighSchool_PR == "-1" | data$College_Score == "-1")
# Indices: 6  19  71  85  88  96 132 183 195 => nine in total

# Remove missing data
data_corr = data[-missing_rows,]

plot(data_corr$HighSchool_PR, data_corr$College_Score,
     main = "High School and College Entrance Exam Scores",
     xlab="High School Entrance Exam Score Percentile Rank (PR)",
     ylab="College Entrance Score")
```



The correlation coefficient is approximately 0.50.

```r
cor(data_corr$HighSchool_PR, data_corr$College_Score)
```

```
## [1] 0.5074473
```