PTT Analysis of Entrance Exam Scores in Taiwan

Christine P. Chai

June 25, 2020

Ongoing work.

Executive Summary

Write something here

1 Introduction

We are 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, we 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, we decided to gather data and create our own analysis. The target audience of this document would be students taking Statistics 101.

2 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

3 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 our friends to report their scores anonymously, this approach can result in a large selection bias. Many of our friends graduated from the same high school and/or college, so we are likely to have similar entrance exam scores.

Hence we retrieved data from the SENIORHIGH (high school)[5] discussion section on PTT[6], the largest terminal-based bulletin board in Taiwan.[7] We assume the data to be more representative (than if we had collected on our 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.

3.1 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
[7] https://en.wikipedia.org/wiki/PTT_Bulletin_Board_System

4 Exploratory Data Analysis

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

4.1 High School Entrance Exam Scores (Percentile Rank)

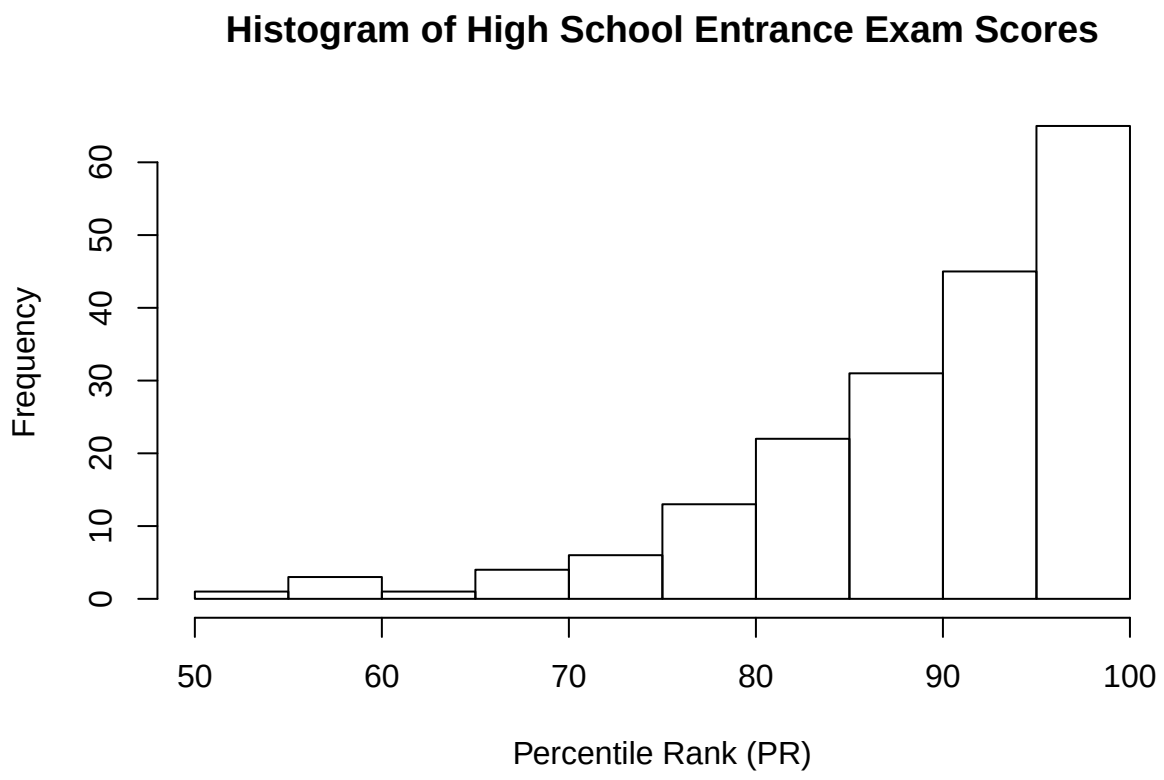
Below shows the descriptive statistics of **HighSchool_PR**, i.e., the percentile rank of high school entrance exam scores. The missing values are removed beforehand. Approximately 75% of the respondents have a percentile rank (PR) at least 85, indicating that most of the respondents scored in the top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



4.2 College Entrance Exam Scores

Similarly, we 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[8] 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.

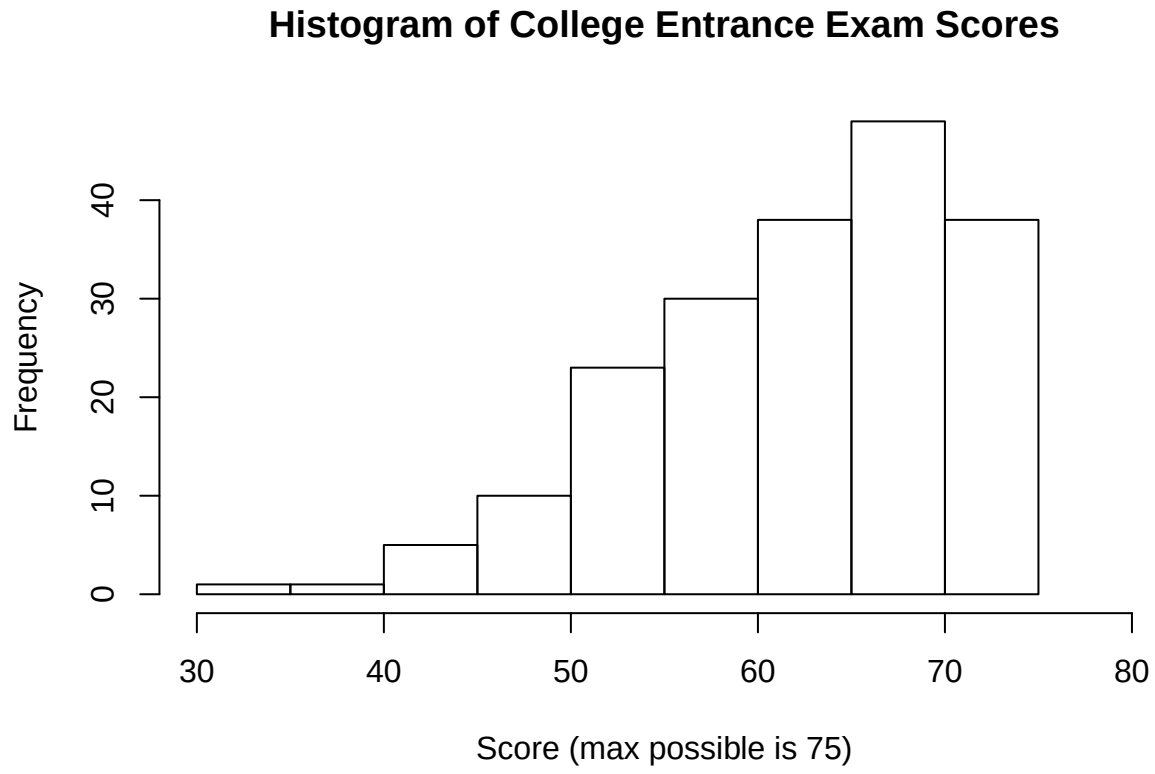
Since PTT contains forums for several prestigious universities in Taiwan, it is no surprise that many users attended these colleges because they scored high on the college entrance exam. Nevertheless, PTT did not limit registration to students of these colleges in the past, so the population of PTT is slightly more diverse. Note that as of 2020, PTT changed their eligibility requirements, and limited new account registrations to only people with an email address from National Taiwan University.[9]

```
# 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))
```



[8] https://bit.ly/3bAYOvO
[9] Screenshot obtained on May 26, 2020: https://imgur.com/33fwrGH

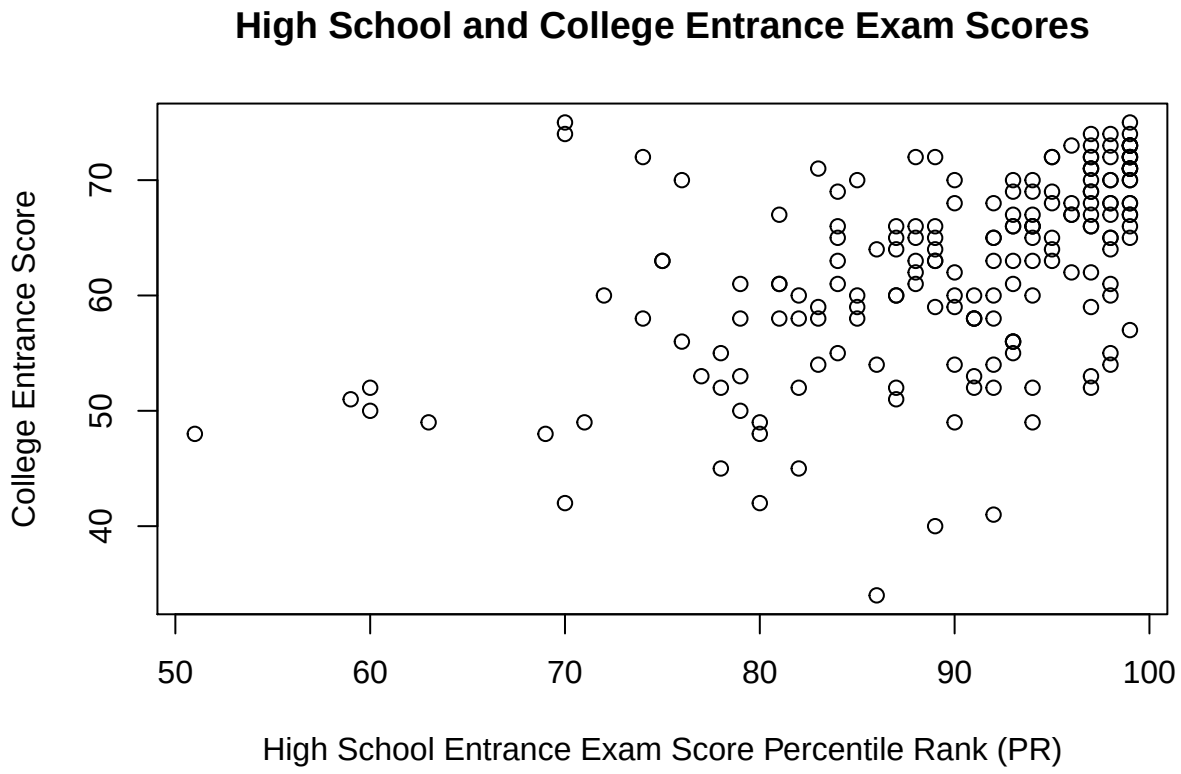
4.3 Bivariate Exploration

Next, we create a bivariate scatterplot of **HighSchool_PR** and **College_Score**, but we have to remove the records with at least one missing score. Just like what we observed in the univariate plots, both variables are largely concentrated towards the maximum possible scores.

```
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 0.507, showing a medium strength of positive association between **HighSchool_PR** and **College_Score**. We can interpret that a better score in the high school entrance exam is likely to lead to a better college entrance exam score, but the relationship is not as strong after **HighSchool_PR** reaches 80.

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

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

5 Statistical Modeling Decisions

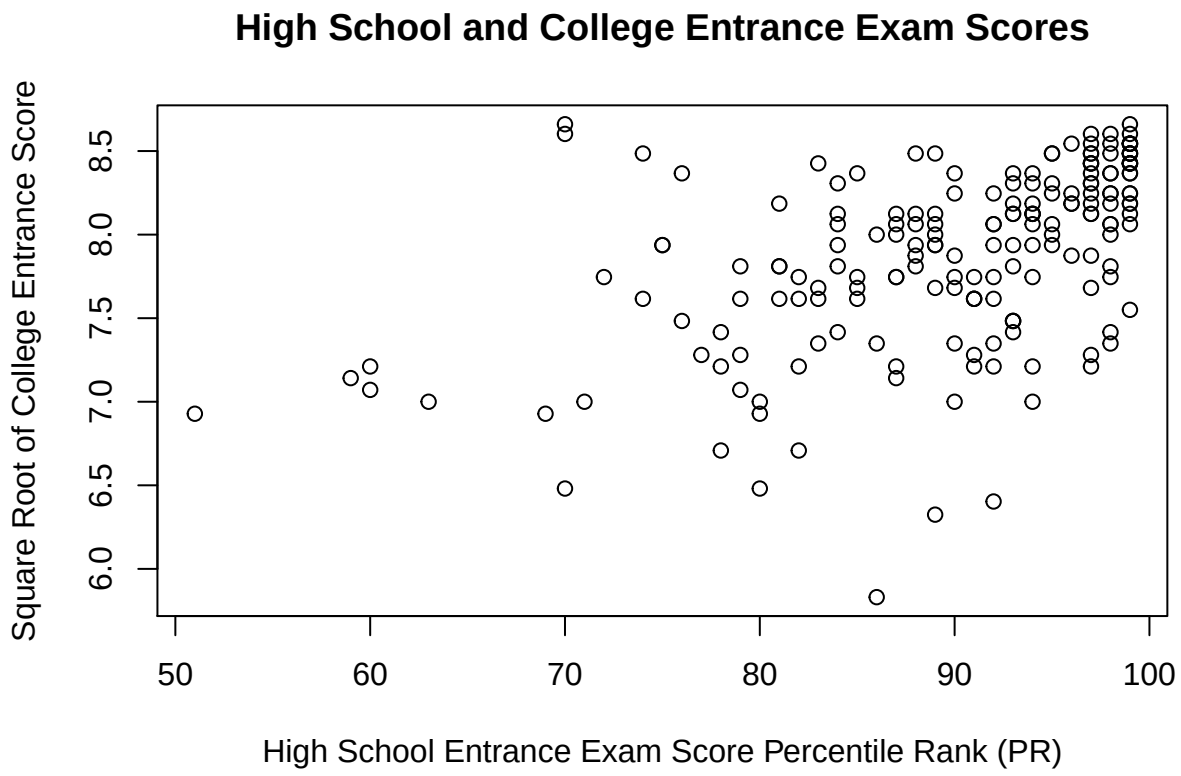
We are going to decide whether we should run a linear regression on **HighSchool_PR** and **College_Score**. If yes, we would implement the model and check the residuals. If no, we need to explore other options in analyzing the data.

5.1 Should we run a linear regression?

It is inappropriate to perform linear regression directly, because the data do not meet the constant variability assumption [2]. In the bivariate exploratory plot, we can see that the variability of **College_Score** increases as **HighSchool_PR** increases. One possible remedy is apply the square root transformation to **College_Score**, in order to reduce the variability. But the scatterplot below shows little to no improvement in variability, and the correlation coefficient even drops from 0.507 to 0.504. Hence we determine that it is not a good idea to run a linear regression model on the whole dataset.

```
# data version: already removed missing data
# data_corr = data[-missing_rows,]

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




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

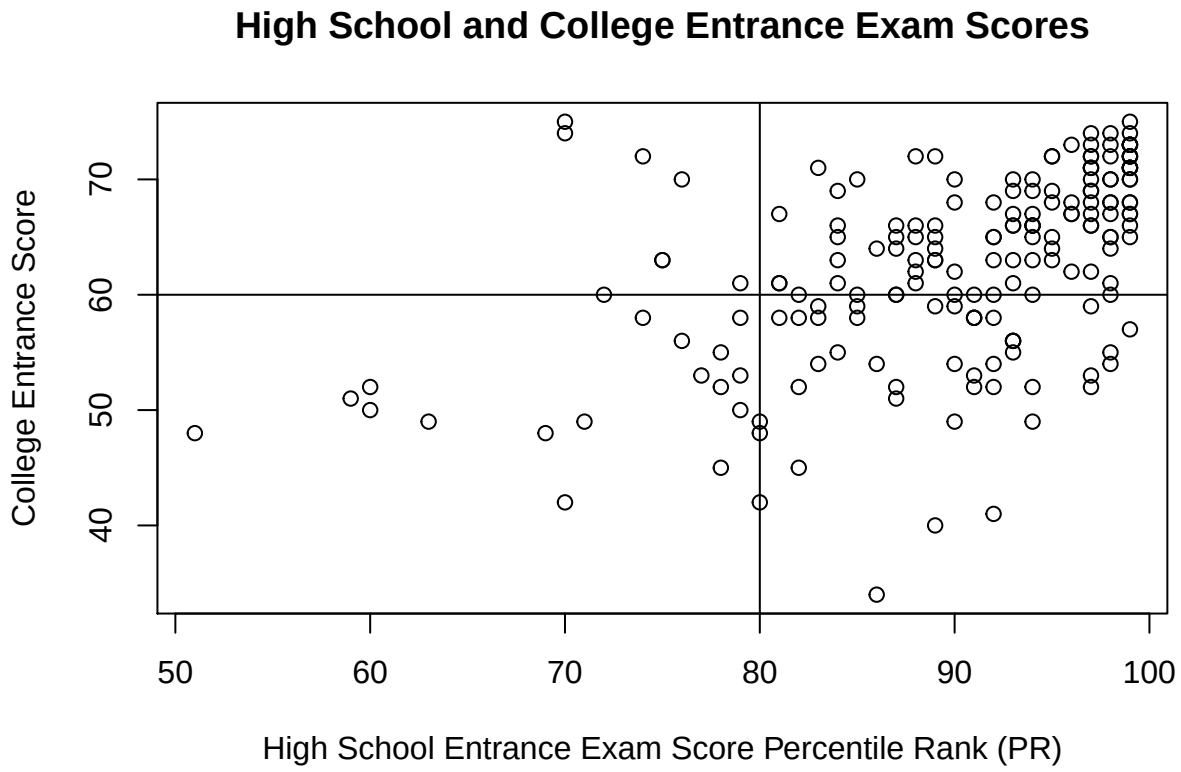
```
## [1] 0.5048132
```

5.2 Segmenting the Data

Instead, we should segment the data and further examine the top scorers in the dataset, i.e., those with **HighSchool_PR** 80 or higher. Most of these respondents have **College_Score** of 60 or higher, but the range of **College_Score** is wide. Here, we add horizontal and vertical lines to clarify the graph.

```
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")

abline(h=60,v=80)
```



We can also create a contingency table for the two indicators:

- **HighSchool_80up**: Indicator of whether **HighSchool_PR** is 80 or higher
- **College_60up**: Indicator of whether **College_Score** is 60 or higher

```
data_corr$HS_80up = data_corr$HighSchool_PR >= 80
data_corr$CS_60up = data_corr$College_Score >=60

contingency = table(data_corr$HS_80up, data_corr$CS_60up,
                    dnn=c("HighSchool_80up","College_60up"))

contingency
```

```
##                College_60up
## HighSchool_80up FALSE TRUE
##           FALSE    17    8
```

```
##           TRUE     43  120
```

Below is the percentage version of the contingency table, and we can see that more than 63.5% of the respondents have both **HighSchool_PR** ≥ 80 and **College_Score** ≥ 60. This is also evidence that the PTT users tend to come from the population who scored well on the high school and college entrance exams.

```
prop.table(contingency)
```

```
##                College_60up
## HighSchool_80up      FALSE       TRUE
##           FALSE 0.09042553 0.04255319
##           TRUE  0.22872340 0.63829787
```

We can also round the percentage table to four decimal places in the ratio, so we will have two decimal places after the integer percentage. For example, 0.4528 becomes 45.28%.

```
round(prop.table(contingency),4)
```

```
##                College_60up
## HighSchool_80up  FALSE   TRUE
##           FALSE 0.0904 0.0426
##           TRUE  0.2287 0.6383
```

5.3 Conditional Probability

Using conditional probability, we can answer this question from the data: If a person scores at least 80 on the high school entrance score percentile rank (PR), how likely is he/she going to obtain a score at least 60 on the college entrance exam?

In mathematical terms, this is equivalent to finding $P(\text{College\_60up is true} \mid \text{HighSchool\_80up is true})$. Recall the conditional probability formula and the Bayes theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

In this data, we have

- $P(\text{HighSchool\_80up is true})$ = # of respondents with **HighSchool_PR** ≥ 80 / all respondents.
- $P(\text{College\_60up is true})$ = # of respondents with **College_Score** ≥ 60 / all respondents.
- $P(\text{HighSchool\_80up is true} \cap \text{College\_60up is true})$
  = # of respondents with **HighSchool_PR** ≥ 80 and **College_Score** ≥ 60 / all respondents.

Plugging the numbers into the equation, we get

$$\begin{aligned}
&P(\text{College_60up is true} \mid \text{HighSchool_80up is true}) \\
&= \frac{P(\text{HighSchool_80up is true} \cap \text{College_60up is true})}{P(\text{HighSchool_80up is true})} \\
&= \frac{\text{\# of respondents with HighSchool_PR} \geq 80 \text{ and College_Score} \geq 60}{\text{\# of respondents with HighSchool_PR} \geq 80} \\
&= \frac{120}{43+120} \approx 0.7362.
\end{aligned}$$

According to this data from PTT, there is a 73.62% chance for a person to score at least 60 on the college entrance exam, given that he/she scored at least 80 on the high school entrance score percentile rank (PR). Note that we use number of respondents rather than percentage to avoid rounding errors.

In comparison, if we do not know anything about the person's high school entrance score percentile rank (PR), we have a probability of 63.82% in observing the person scoring at least 60 on the college entrance exam. There is an increase of 9.80% in probability after we learn information about his/her high school entrance exam score.

$$\begin{aligned} P(\text{College\_60up is true}) &= \#\text{ of respondents with College\_Score} \geq 60 \text{ / all respondents} \\ &= \frac{120}{188} \approx 0.6382. \end{aligned}$$

```
nrow(data_corr) # number of all respondents without missing data
```

```
## [1] 188
```

Remark: Conditional probability is the foundation of Bayesian statistics, which updates the existing probabilities given the new data. For the interested readers, we recommend the book *An Introduction to Bayesian Thinking: A Companion to the Statistics with* R *Course* [1] as a starting point to learn Bayesian statistics. It is the supplementary materials for the Bayesian statistics course on Coursera from Duke University[10].

6 A Closer Look at the Top Scorers

We are going to take a closer look at the top scorers, given the observation in Section 5.2. In Taiwan's education system, the top tier of high schools and colleges can be further segmented. The top of the top tier can be very different than the bottom of top tier. Therefore, we consider the following subcategories:

- **HighSchool_PR** ranges: 80-89, 90-94, 95-99
- **College_Score** ranges: 60-64, 65-69, 70-75

6.1 Data Consistency

Before we start the analysis, we need to ensure consistency in the data. For instance, there are 191 records of valid high school entrance exam scores in the data. But if we consider only the ones with a valid college entrance exam score, the number of available records drops to 188. Although which version we use does not matter much when we look at the univariate distribution, this is going to be problematic when we combine the univariate analysis with the bivariate analysis. Thus, we should use only the 188 records whose college entrance exam scores are also valid.

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

```
## [1] 191
```

```
# Consider only the records with both valid HighSchool_PR and College_Score
length(data_corr$HighSchool_PR)
```

```
## [1] 188
```

The same data consistency requirement also applies to the college entrance scores. There are 194 records of valid high school entrance exam scores in the data, but only 188 of them also have corresponding valid high school entrance exam scores. Accordingly, we should use only the 188 records whose high school entrance exam scores are also valid.

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

[10] https://www.coursera.org/learn/bayesian

```
## [1] 194
```

```
# Consider only the records with both valid HighSchool_PR and College_Score
length(data_corr$College_Score)
```

```
## [1] 188
```

6.2 High School Entrance Exam Scores (Percentile Rank)

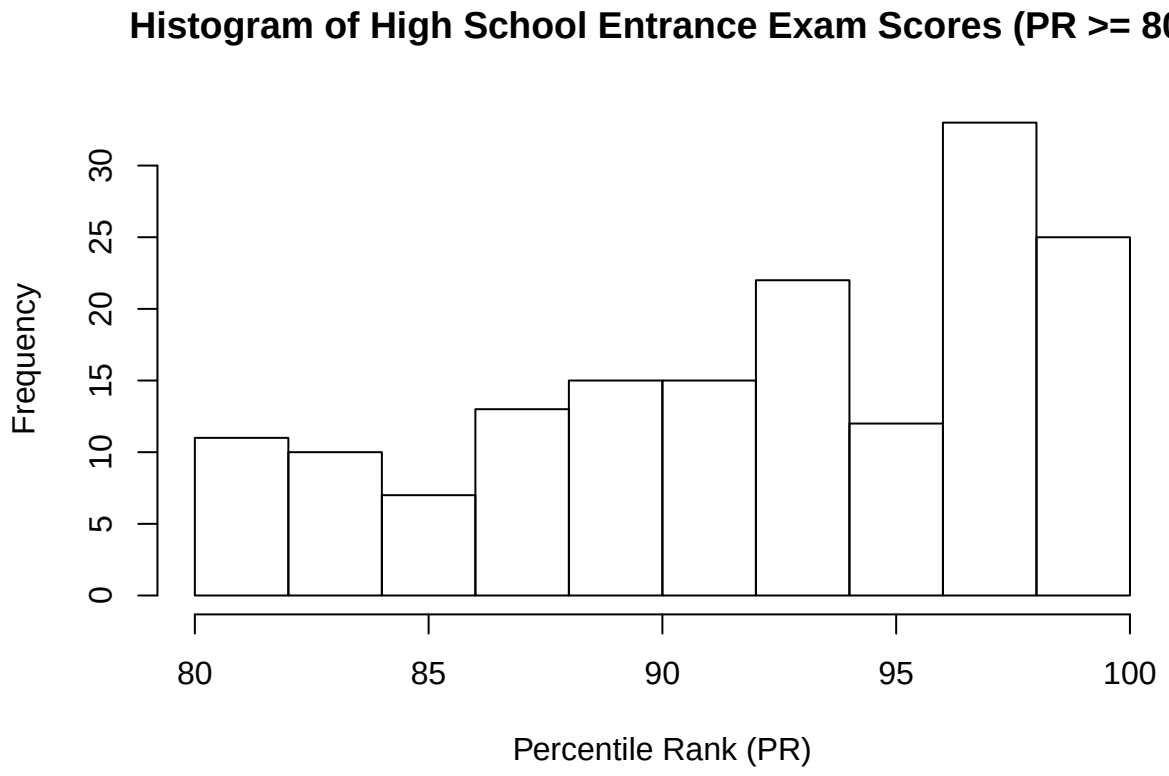
We use the R function `table` to show the frequency of each **HighSchool_PR** value that is at least 80. In the table below, the first row is the PR (percentile rank), and the second row is the counts. Although we truncated the **HighSchool_PR** to 80 and above, the distribution is still left-skewed. The **HighSchool_PR** 99 has the highest counts, followed by **HighSchool_PR** 97 and **HighSchool_PR** 98.

```
HS_PR_seg = data_corr$HighSchool_PR[which(data_corr$HS_80up == TRUE)]
table(HS_PR_seg)
```

```
## HS_PR_seg
## 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
##  3  4  4  4  6  4  3  7  6  8  7  6  9 11 11  7  5 18 15 25
```

Or if you prefer a histogram, we can also create one.

```
hist(HS_PR_seg, xlab="Percentile Rank (PR)",
     main="Histogram of High School Entrance Exam Scores (PR >= 80)")
```



Unfinished below

- **HighSchool_PR** ranges: 80-89, 90-94, 95-99

Therefore, we create the breakdown of the **HighSchool_PR** ranges: 80-89, 90-94, 95-99. There are 49 records in the 80-89 range, 44 records in 90-94, and 70 records in 95-99. We divided the 90-99 range into 90-94 and 95-99, but the number of **HighSchool_PR** records in the 95-99 range is still higher than any of the other two categories.

However, it is not a good idea to further divide the 95-99 range into 95-97 and 98-99, due to the lack of geographic information. In Taipei, the high school enrollment is extremely competitive. Students with **HighSchool_PR** 95 and those with **HighSchool_PR** 99 would get admitted to high schools of different ranking[11]. But in other parts of Taiwan, most students with **HighSchool_PR** at least 95 would already qualify for the top local high school, and some rural parts even require a lower **HighSchool_PR** to get into the county's top high school[12].

```
HS80to89 = length(which(HS_PR_seg >= 80 & HS_PR_seg <= 89))
HS90to94 = length(which(HS_PR_seg >= 90 & HS_PR_seg <= 94))
HS95to99 = length(which(HS_PR_seg >= 95 & HS_PR_seg <= 99))
```

```
print(paste("HighSchool_PR 80-89:",HS80to89))
```

```
## [1] "HighSchool_PR 80-89: 49"
```

```
print(paste("HighSchool_PR 90-94:",HS90to94))
```

```
## [1] "HighSchool_PR 90-94: 44"
```

```
print(paste("HighSchool_PR 95-99:",HS95to99))
```

```
## [1] "HighSchool_PR 95-99: 70"
```

6.3 College Entrance Exam Scores

- **College_Score** ranges: 60-64, 65-69, 70-75

Write something here

```
CS_Score_seg = data_corr$College_Score[which(data_corr$CS_60up == TRUE)]
table(CS_Score_seg)
```

```
## CS_Score_seg
## 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
## 10  7  4 10  5 11 12  9  9  6 10  9 12  8  4  2
```

6.4 Bivariate Exploration

- **HighSchool_PR** ranges: 80-89, 90-94, 95-99
- **College_Score** ranges: 60-64, 65-69, 70-75

Write something here

7 More Sections coming up

More stuff

[11] https://w199584.pixnet.net/blog/post/28321887
[12] http://www.edtung.com/TopNews/NewsContent.aspx?type=4&no=1278

8 Final: Personal Remarks

Even more stuff

References

[1] Merlise Clyde, Mine Cetinkaya-Rundel, Colin Rundel, David Banks, Christine Chai, and Lizzy Huang. *An Introduction to Bayesian Thinking: A Companion to the Statistics with* R *Course*. GitHub, 2020. Available from: https://statswithr.github.io/book/.

[2] David M Diez, Mine Cetinkaya-Rundel, and Christopher D Barr. *OpenIntro Statistics*. OpenIntro, 4th edition, 2019. Available from: https://www.openintro.org/book/os/.