SAT Scores in NYC: STAT 3080 Project Part 3

David Bass

Question

Does geography determine SAT score in New York City? In this project, I will examine the relationship between the borough and average SAT score of accredited high schools in NYC in the 2014-15 school year.

Data Description

My data, obtained from Kaggle.com and collected by NYC Open Data, describes the population of all 435 accredited high schools in New York City in New York City. (1) I will be using the 374 rows (each row represents an observation of 1 school) for which the data set had complete data on location and average SAT score. (2)

Data Relevance

My data is appropriate for answering my quesiton because it allows me to group schools based on borough and categories of average SAT score ranges. I will be able to make a 2d matrix of counts with borough as one of the axes and SAT score intervals as the other, which will allow for the justified application of a chi-squared test for independence between borough and average SAT score.

Generalization

My data represents the population of accredited NYC high schools in the 2014-15 schools year, so the outcome of my chi-squared test for independence will be immediately applicable to that population. Assuming that no major changes in NYC's demographics or education sytem have occurred in the last 5 years, my results will be applicable to present-day NYC schools. By comparing socioeconomic and demographic characteristics of NYC's boroughs to those of other urban areas in the US, my results could be extrapolated to examine educational disparities among other American urban localities.

Test Appropriateness

The chi-squared test for independence is appropriate to answer my question because both geography and SAT score can be described as categorical variables, with each borough as a geographic category, and 3 categories of "low" (<1200 points), "medium" (>=1200 points, but <1400 points), and "high" (>= 1400 points) representing SAT score categories. If the chi-squared test for independence rejects its null hypothesis, then it a relationship between geography and SAT score because the data cannot be assumed to be independent.

Test Selection

The assumption of the test that the data is sampled randomly from the population is satisfied because the data represents the whole population. With the removal of Staten Island (only 10 schools, 2.67% of total), the assumption that all cells contain values greater than 0 is satisfied, and the assumption that all expected cell values be greater than 5 is easily satisfied. (3)

Data Preparation

```
setwd("C:/Users/David/Desktop/Programs/STAT 3080/Project")
scores_raw <- read.csv("scores.csv")
scores <- scores_raw[is.na(scores_raw$Average.Score..SAT.Writing.)==FALSE &
                     scores_raw$School.Name!="Forest Hills High School",
                     c("Borough", "Average.Score..SAT.Total.")]
names(scores) <- c("borough", "total")
m <- cbind(c(dim(scores[scores$borough == "Manhattan" & scores$total < 1200,])[1],
             dim(scores[scores$borough == "Manhattan" & 1200 <= scores$total &
                        scores$total < 1400,])[1],
             dim(scores[scores$borough == "Manhattan" & scores$total >= 1400,])[1]),
           c(dim(scores[scores$borough == "Bronx" & scores$total < 1200,])[1],
             dim(scores[scores$borough == "Bronx" & 1200 <= scores$total &
                        scores$total < 1400,])[1],
             dim(scores[scores$borough == "Bronx" & scores$total >= 1400,])[1]),
           c(dim(scores[scores$borough == "Queens" & scores$total < 1200,])[1],
             dim(scores[scores$borough == "Queens" & 1200 <= scores$total &
                          scores$total < 1400,])[1],
             dim(scores[scores$borough == "Queens" & scores$total >= 1400,])[1]),
           c(dim(scores[scores$borough == "Brooklyn" & scores$total < 1200,])[1],
             dim(scores[scores$borough == "Brooklyn" & 1200 <= scores$total &
                          scores$total < 1400,])[1],
             dim(scores[scores$borough == "Brooklyn" & scores$total >= 1400,])[1]))
m
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   30   53   12   59
## [2,]   36   40   35   37
## [3,]   23    5   21   13
```

Performing the Test

```
chisq.test(m, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  m
## X-squared = 42.878, df = 6, p-value = 1.233e-07
```

Test Conclusions

The chi-squared test for independence has a test-statistic value of 42.88, which corresponds to a p-value of .00000012 at (3-1)(4-1)=6 degrees of freedom. Thus, we can confidently reject the null hypothesis that borough and average SAT score level are independent at a 95% confidence level and conclude that there is a statistically significant relationship between borough and SAT score level because our p-value is less than 0.05. In less statistical terms: Yes, average SAT scores by high school vary significantly among the boroughs of NYC. These results can be generalized to other urban areas in the US by comparing the characteristics of each borough to other localities with confidence in the assumption that a locality's characteristics have a strong influence on the SAT scores of students attending their schools. (4)

References

1. https://opendata.cityofnewyork.us/
2. https://www.kaggle.com/nycopendata/high-schools
3. https://online.stat.psu.edu/stat200/book/export/html/230
4. https://www.dummies.com/education/math/statistics/generalizing-statistical-results-to-the-entire-population/