High School and Beyond Data Analysis Project

Katie Tucker

9/25/2018

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
library(ggplot2)
library(dplyr)
hsb2 <- read.delim("/Users/katherinetucker/Documents/Katie_Work_3/homework/Math130/hsb2.txt",
                   header = TRUE, sep="\t")
```

Introduction

The high school and beyond data set is a subset of the high school and beyond longitudinal study which was collected by surveying students in high school and recoding their scores on tests. The same group of students was surveyed again after two, four, six and twelve years had passed since the first survey. The subset only covers one of these surveys for 200 of the 58,000 high school students that participated in the study. Of the information included in the high school and beyond data set this project will specifically examine the students test scores in science and the type of program that they are enroled in.

Description of Variables

Type of Program

The type of school program is described by the variable named `prog` which has three categories: academic, general, and vocational.

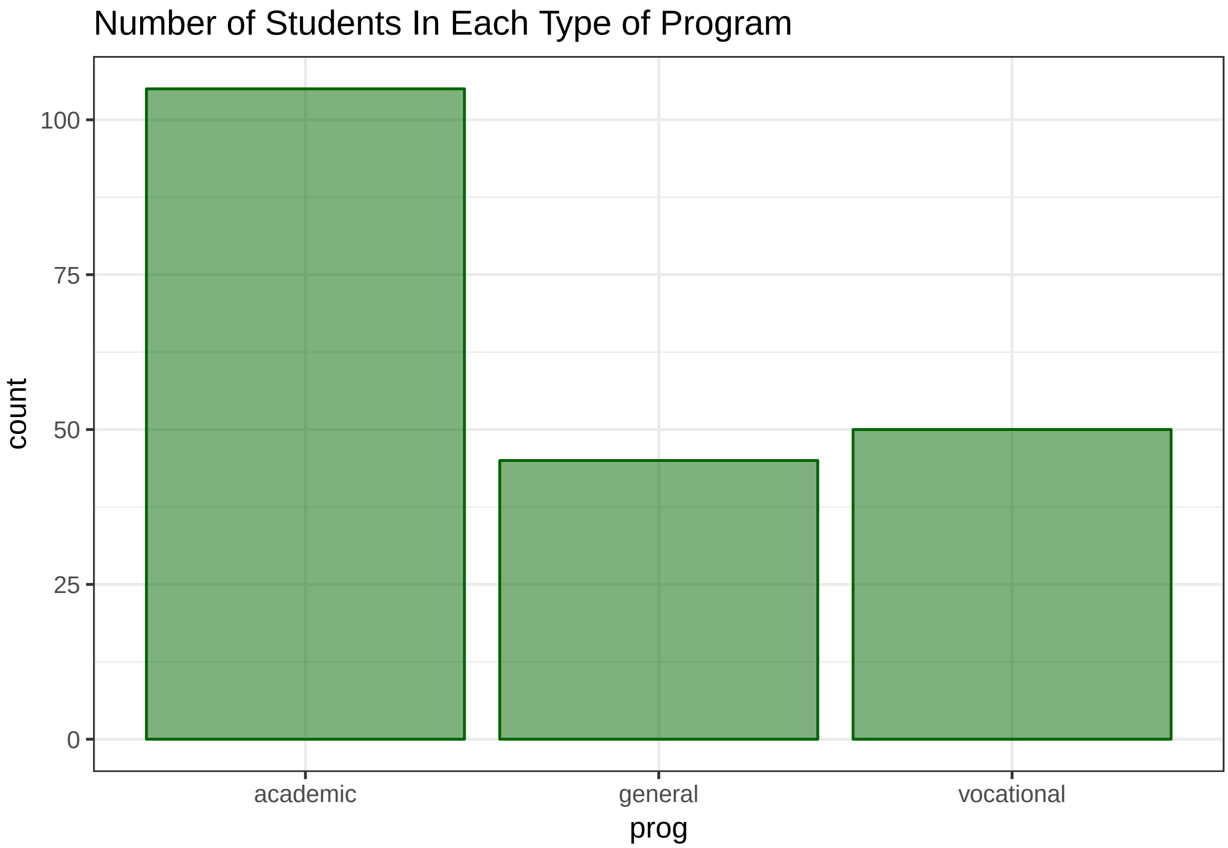
```
#Number of people surveyed in each type of program.
table(hsb2$prog)
```

```
##
##   academic    general vocational
##        105         45         50
```

```
#Proportion of total people surveyed in each type of program.
prop.table(table(hsb2$prog))
```

```
##
##   academic    general vocational
##      0.525      0.225      0.250
```

```
ggplot(hsb2, aes(x=prog)) +
  geom_bar(col="dark green", fill = "dark green", alpha = .5) +
  ggtitle("Number of Students In Each Type of Program") +
  theme_bw()
```



The two hundred students surveyed in the data set are all part of one of three programs. One hundred and five students (52.5%) are enroled in academic programs. Fourty five students (22.5%) are enroled in general programs and fifty students (25%) are enroled in vocational programs.

Science Scores

Scores in standardized tests, including science tests, were collected along with survey responses.

Summary Statistics

```r
summary(hsb2$science)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   44.00   53.00   51.85   58.00   74.00
```

```r
#Variance
var(hsb2$science)
```

```
## [1] 98.02764
```

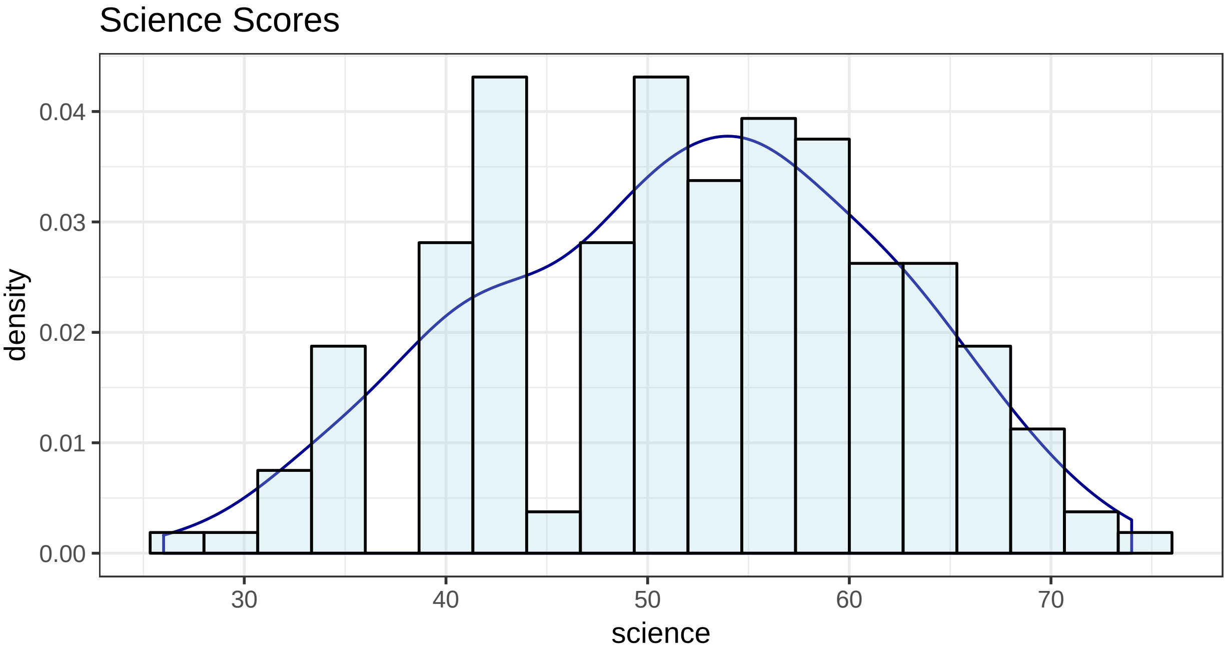
```r
#Standard Deviation
sd(hsb2$science)
```

```
## [1] 9.900891
```

The science test scores range from twenty six to seventy four with a median score of 53 and a mean score of 51.85. The variance of the scores is 98.0 and the standard deviation is 9.9. The first quartile is 44 and the third quartile is 58.

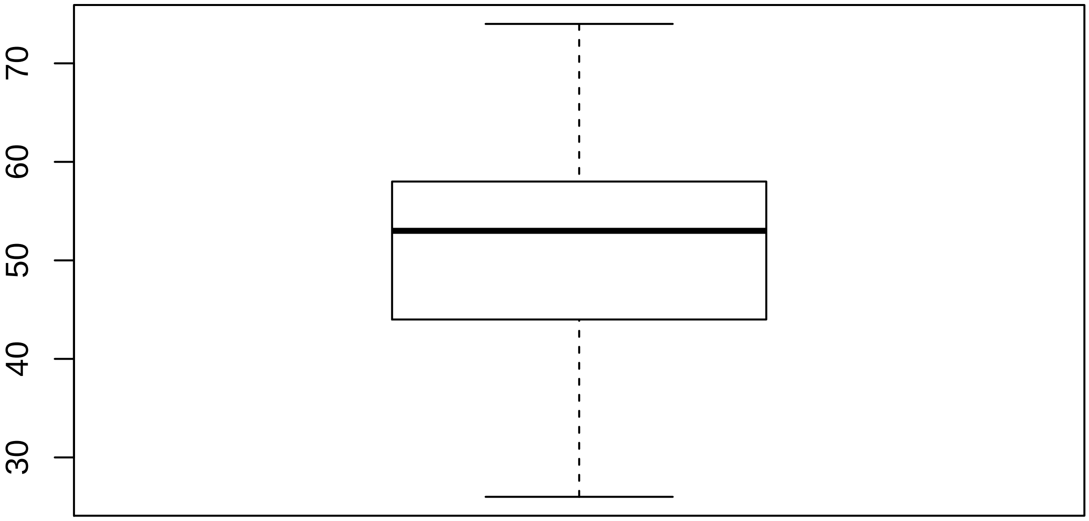
Graphs of Science Scores

```
ggplot(hsb2, aes(x=science)) +
  geom_density(col= "dark blue", alpha = 1) +
  geom_histogram(aes(y=..density..), col= "black", bins = 19, fill = "light blue", alpha = .3) +
  ggtitle("Science Scores") +
  theme_bw()
```



The histogram and density plot for the science scores is slightly skewed to the right and unimodal despite a slight concentration of scores slightly to the left of the main group of scores.

```
boxplot(hsb2$science)
```

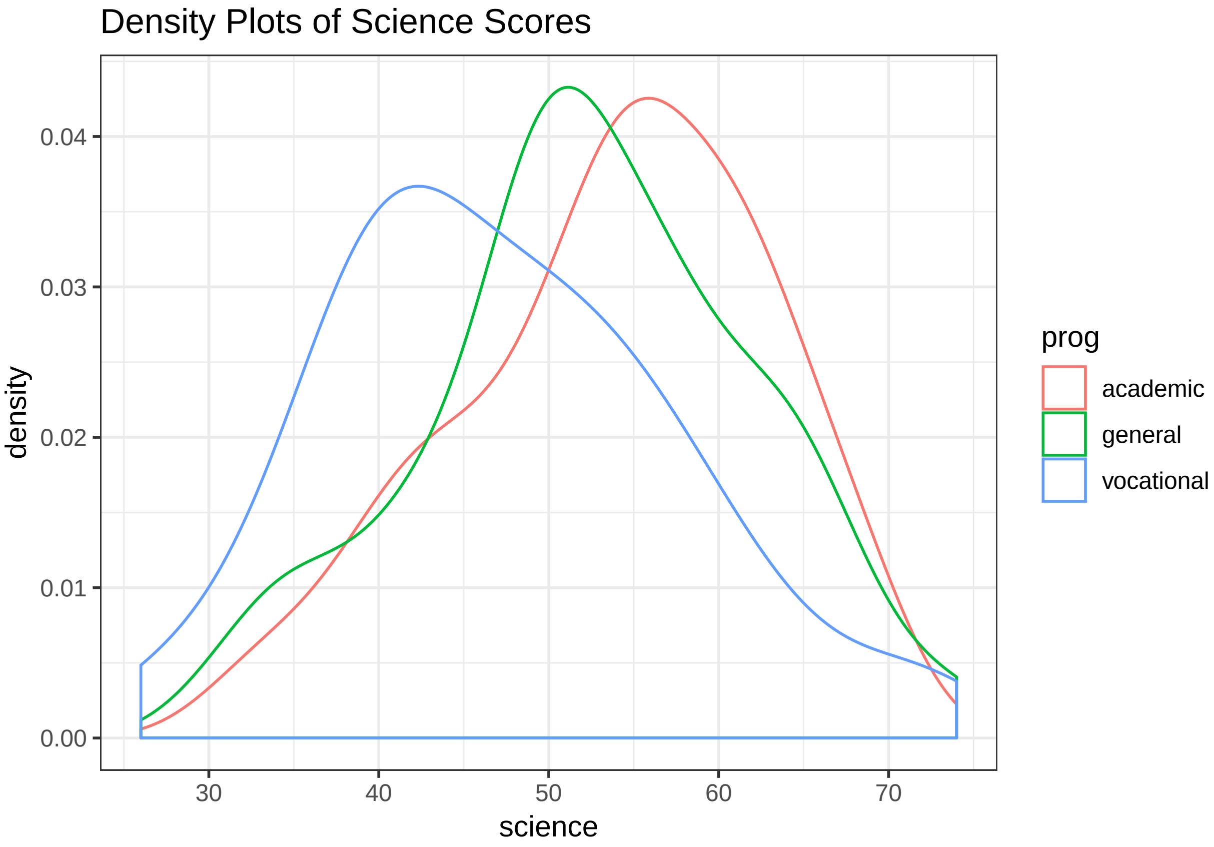


The box plot of the science scores shows that there are no outliers or suspected outliers in the data. It also shows that a greater number of scores are below the overall median score than above it.

Bivariate Comparison

Density Plots

```
ggplot(hsb2, aes(x=science, col = prog)) +
  geom_density() +
  ggtitle("Density Plots of Science Scores") +
  theme_bw()
```



The overlain density plots of each academic program show that the science scores for academic and general programs are skewed to the right and the science scores for the vocational programs are skewed toward the left. All three density plots are unimodal.

Summary Statistics

```
academic_sci <- filter(hsb2, prog == "academic")

summary(academic_sci$science)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    31.0    48.0    55.0    53.8    61.0    69.0
```

```
#Variance
var(academic_sci$science)
```

```
## [1] 83.31538
```

```
#Standard Deviation
sd(academic_sci$science)
```

```
## [1] 9.127726
```

The science test scores for the academic program students range from the minimum score of thiry one to the maximum score of sixty nine. The first quartile is fourty eight and the third quartile is sixty one. The mean score is 53.8 and the median score is fifty five. The variance is 83.3 and the standard devation is 9.1.

```
general_sci <- filter(hsb2, prog == "general")

summary(general_sci$science)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   31.00   47.00   53.00   52.44   58.00   74.00
```

```
#Variance
var(general_sci$science)
```

```
## [1] 93.70707
```

```
#Standard Deviation
sd(general_sci$science)
```

```
## [1] 9.680241
```

The science scores for the general program students range form the minimum of thirty one to the maximum of seventy four. The first quartile is fourty sevena and the third quartile is fifty eight. The mean score is 52.4 and the median score is fifty three. The variance is 93.7 and the standard deviation is 9.7.

```
vocational_sci <- filter(hsb2, prog == "vocational")

summary(vocational_sci$science)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   26.00   39.00   47.00   47.22   53.75   72.00
```

```
#Variance
var(vocational_sci$science)
```

```
## [1] 106.7873
```

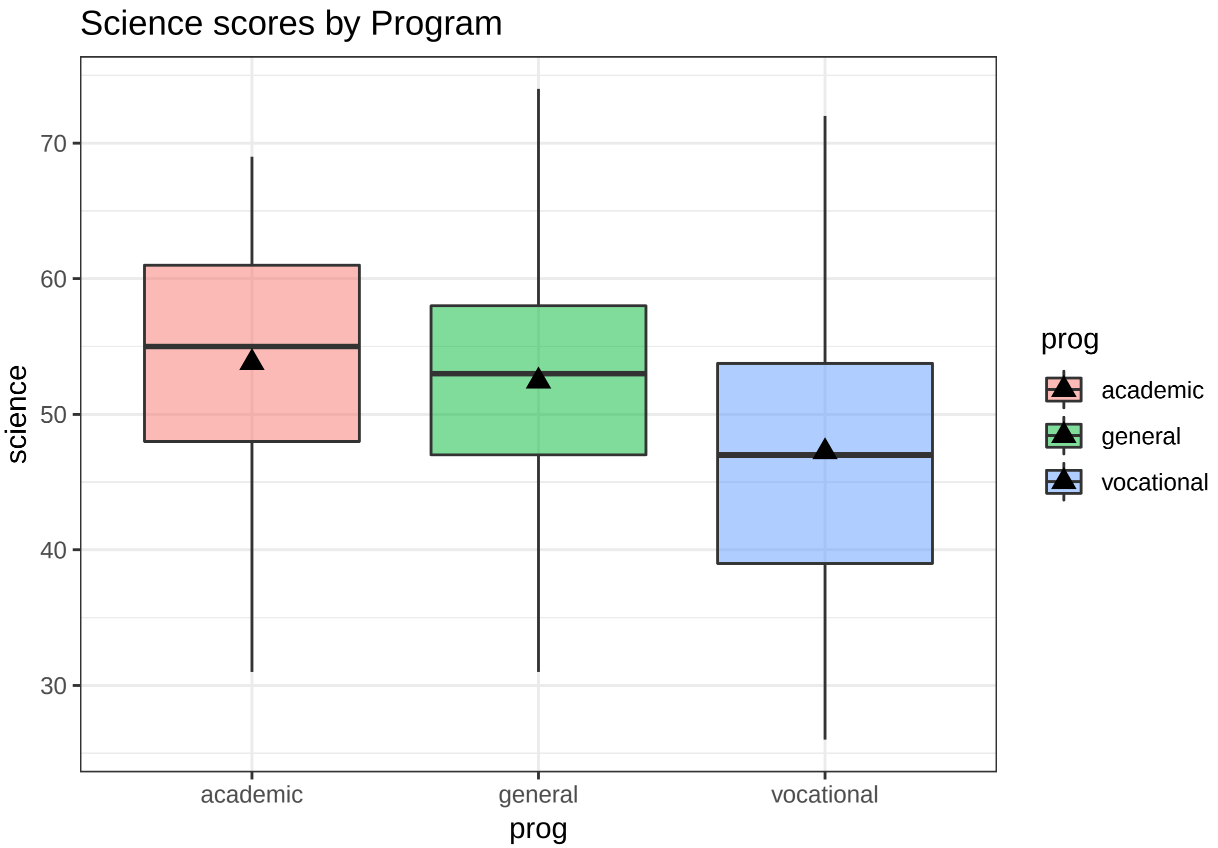
```
#Standard Deviation
sd(vocational_sci$science)
```

```
## [1] 10.3338
```

The science scores for the vocational program students frange from a minimum score of twenty six to a maximum of seventy two. The first quartile is thirty nine and the third quartile is 53.75. The mean score is 47.22 and the median score is fourty seven. The variance is 106.8 and the standard deviation is 10.3.

Grouped Boxplot

```
ggplot(hsb2, aes(x= prog, y=science, fill = prog)) +
  geom_boxplot(alpha = .5) +
  stat_summary(fun.y="mean", geom="point", size=3, pch=17,
    position=position_dodge(width=0.75)) +
  ggtitle("Science scores by Program") +
  theme_bw()
```



The grouped box plot shows that the academic program has the highest median and mean score of all of the programs. The highest score of all of the programs occurs in the general program and the lowest score of all of the programs occurs in the vocational program. The interqartile range is the largest for the vocational program followed by the academic program with the general program having the smallest interquartile range. The scores encompased by the interquartile range are the highest for the academic programs, in the middle for the general programs and the lowest for the vocational programs.