APSTA-GE 2352 Project 1

Kenny Mai

11/2/2020

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
LatinoEd <- read_csv("~/R/2352-Stat-Computing/LatinoEd.csv")
```

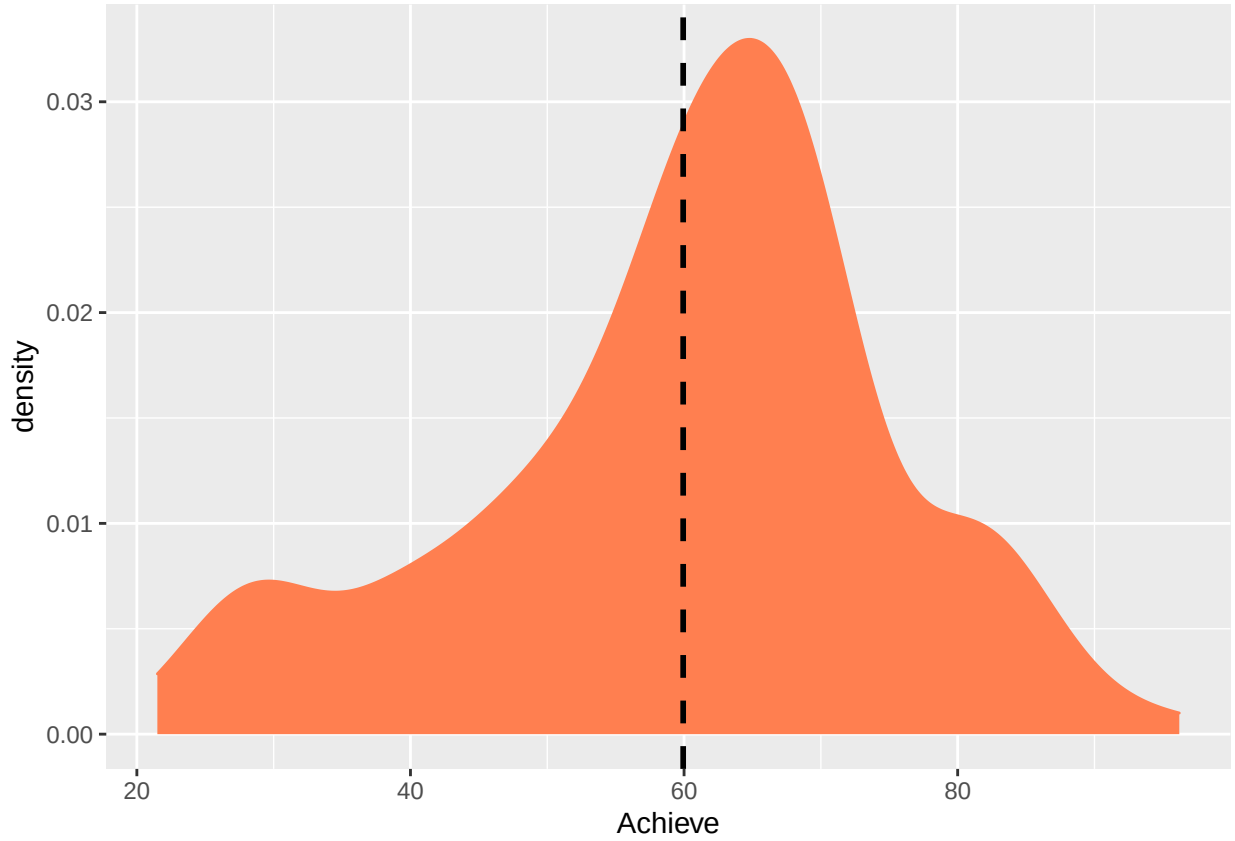
```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Achieve = col_double(),
##   ImmYear = col_double(),
##   ImmAge = col_double(),
##   English = col_double(),
##   Mex = col_double()
## )
```

Problem 4.1

Data on U.S. citizens were collected in 2000 by the U.S. Census Bureau. The data set LatinoEd.csv—a subset of the Census data—contains data for a sample of Latino immigrants who were naturalized U.S. citizens living in Los Angeles. The variable Achieve in this data set provides a measure of educational achievement (see the codebook for more information regarding the interpretation of this variable).

a) Construct a plot of the kernel density estimate for the marginal distribution of the educational achievement variable. Discuss all interesting features of this plot.

```
ggplot() +
  geom_density(data=LatinoEd,aes(x=Achieve),color="coral",fill="coral") +
  geom_vline(data=LatinoEd,aes(xintercept=mean(LatinoEd$Achieve)),linetype="dashed",color="black",size=
```

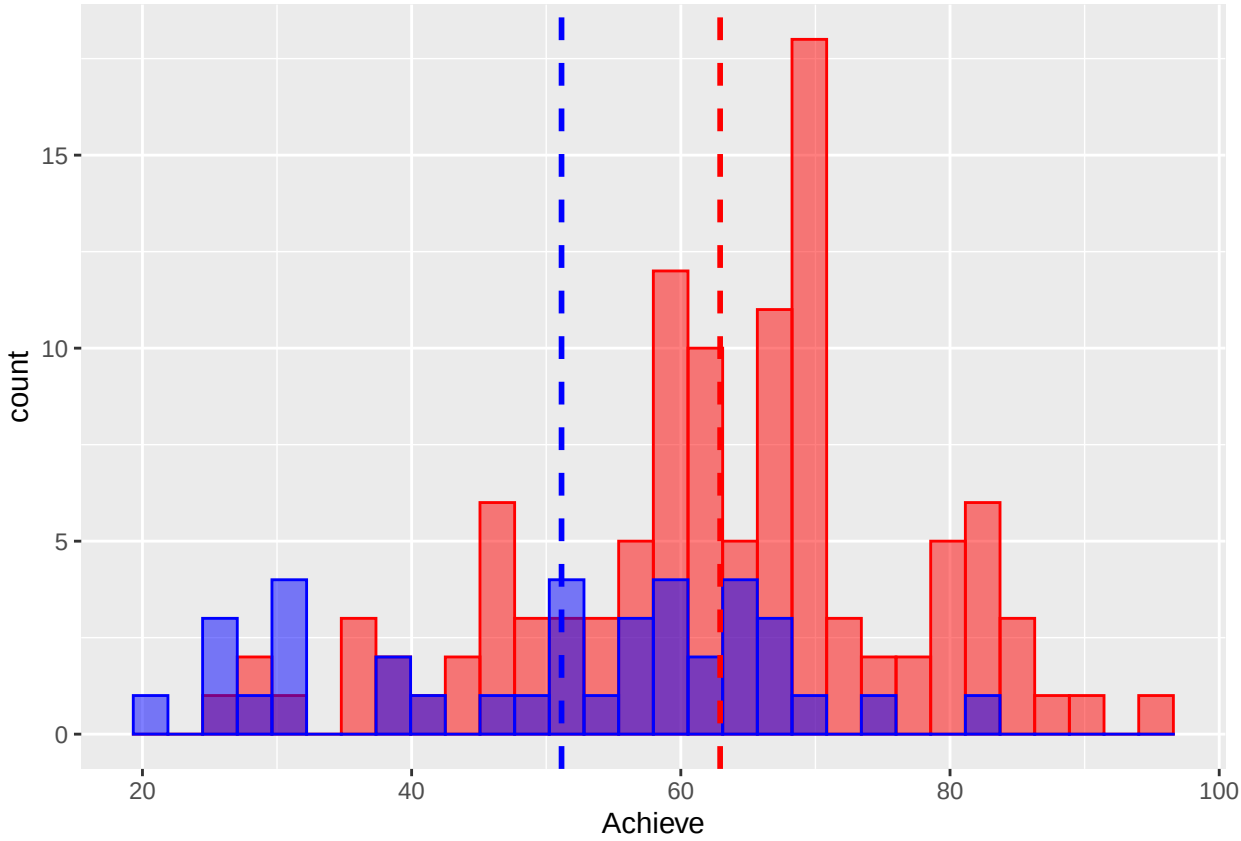


The key interesting feature of this plot is that the mean of the variable Achieve occurs not at the peak of the density shape, but slightly below it, at value 60. The distribution of this variable looks to be close to normal, albiet slightly negatively skewed.

b) Examine plots of the variable Achieve conditioned on the variable English to compare the educational achievement of Latino immigrants who are fluent in English and those who are not. Create a single publishable display that you believe is the best visual representation of the results of this analysis. In constructing this display, think about the substantive points you want to make and create a graph that best allows you to highlight these conclusions. Write a brief paragraph explaining why you chose to construct your graph the way you did and how it helps answer the research question of interest.

```
eng <- LatinoEd %>% filter(English==1)
noeng <- LatinoEd %>% filter(English==0)
ggplot() +
  geom_histogram(data=eng,aes(x=Achieve),color="red",fill="red",alpha=0.5) +
  geom_histogram(data=noeng,aes(x=Achieve),color="blue",fill="blue",alpha=0.5) +
  geom_vline(data=eng, aes(xintercept=mean(eng$Achieve)),color="red",linetype="dashed",size=1) +
  geom_vline(data=eng, aes(xintercept=mean(noeng$Achieve)),color="blue",linetype="dashed",size=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Here, overlapping histograms were used to show conditional distributions of the variable Achieve conditioned on the variable English. Namely, red bars and a red dashed line were used to show the distribution of Achieve scores for English speaking chidlren and the mean value, and blue bars and a blue dashed line were used to show non-English speaking children. It can be seen that when the variable Achieve is deconstructed, children who speak English not only had higher educational outcomes, but those outcomes occurred at greater frequency than those children who could not speak English. The mean educational outcome for English fluent children is now higher than compared to the mean of the marginal distribution before conditioning (mean value around 60), and the mean educational outcome for children not fluent in English is lower.

c) Compute appropriate numerical summaries of achievement conditioned on English fluency. Use these summaries (along with evidence culled from your plot) to provide a full comparison of the distributions of achievement scores for immigrants who are fluent in English and those immigrants who are not fluent in English. Be sure to make comparisons between the measures of center and variation in the distributions. Use the problem context to help you write your answer. Be selective in what you report, remembering that you want to be succinct yet thorough.

```
stat_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
LatinoEd %>% group_by(English) %>% summarize(n=n(),mean=mean(Achieve),mode=stat_mode(Achieve))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 4
##   English     n  mean  mode
##     <dbl> <int> <dbl> <dbl>
## 1       0    38  51.1  63.7
## 2       1   112  62.9  70.2
```

If comparing the visual densities between the two groups of children, it can be seen that the statistical modes are relatively close together, occurring at $x = 63.7$ for children not fluent in English and $x = 70.2$ for children who are not. However, when calculating for the statistical mean, it can be seen that the differences between the two groups are wider, with a mean of 51.1 for non-fluent children and 62.9 for fluent children. This is supported by the overlapping histograms. This answers the following research question: how do educational outcomes for Latino immigrant children who are fluent in English compare with those who are not fluent in English?