Final Project

Kevin Flores

April 6, 2019

knitr::opts\_chunk$set(echo = TRUE)
hsb2 <- read.delim("C:/Users/sick riffs/Documents/MATH130/data/hsb2.txt", sep="\t")
library(ggplot2) ; library(dplyr)

## Warning: package 'ggplot2' was built under R version 3.5.3

##
## 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

# 1. Introduction of the Data

This data set was collected to study the educational development of students starting from high school. A sample of students was used and four follow up surveys were also done. For my project, I wanted to see how the socioeconomic status of the students affects their math scores.

# 2. Univariate

## Socioeconomic Standing

table(hsb2$ses)

##
## high low middle
## 58 47 95

The table shows the socioeconomic status of the paricipants. There are three groups: low, middle, and high.

prop.table(table(hsb2$ses))

##
## high low middle
## 0.290 0.235 0.475

ggplot(hsb2, aes(x= ses)) + geom\_bar(fill= "light blue") + ggtitle("Socioeconomic status of the participants" )



According to the graph the group with the most paricipants belonging to it are the middle with 95 participants or about 48% of the total paticipants in the study. The group the lowest are the low class with 47 participants in it or about 24%. The high class has 58 participants belonging to it or about 29% of the total participants.

## Math Scores

summary(hsb2$math)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 33.00 45.00 52.00 52.65 59.00 75.00

This table shows the min, 1st quartile, median, mean, 3rd quartile, and max for the math scores recorded in the study.

ggplot(hsb2, aes(x= math)) + geom\_density(col="blue") + geom\_histogram(aes(y=..density..),color="black", bins= 20, alpha=.5, fill="light blue") + ggtitle("Histogram of Math Scores")



Histogram shows the distribution of the math scores. The shape of the histogram seems to be skewed to the right and slightly bimodal. The math scores of about 43 and 53 appear to have the highest frequencies.

# Bivariate

## Socioeconomic Standing vs. Math Scores

low <- filter(hsb2, ses == "low")
middle <- filter(hsb2, ses == "middle")
high <- filter(hsb2, ses == "high")

summary(low$math)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 39.00 43.00 46.00 49.17 53.00 72.00

summary(middle$math)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 33.00 45.50 52.00 52.21 57.00 75.00

summary(high$math)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 38.00 50.25 57.00 56.17 62.75 71.00

ggplot(hsb2, aes(x= ses, y= math, fill= ses)) + geom\_violin(alpha=.1) + geom\_boxplot(alpha=.6, width=.4) + stat\_summary(fun.y= "mean", geom= "point", size=3, pch= 18, color="red") + ggtitle("Boxplots of Math Scores")



The summary tables and boxplots all show the min, 1st qu., median, mean, 3rd qu., and max for low, middle, and high variables. The median for the low, middle, and high are 46, 52, and 57 respectively. The mean for the same classes are 49.17 for low, 52.21 for middle, and 56.17 for high. Participants in the low class scored the lowest of the three classes while the middle and high classes had math scores that are similar. This indicates that the socioeconomic status of a student influences their scores in math tests.