07 Comparison of Means Lab (26 points)

# Today’s Lesson

This week’s lab uses data from last semester’s DATA 205 course.

We start by loading our libraries. Today we’re using an Excel dataset I created from my gradebook. To process it easily, we need to include the package readxl in addition to tidyverse.

# Exercise 1: Exploratory Data Analysis (10 points)

The dataset contains two variables. Both are ratio level of measurement:

* test1 – students’ scores for test 1
* perusall – students’ average score on Perusall up to the test date

#reading in the dataset  
testdata <- read\_excel("/cloud/project/test1\_corrected.xlsx")

#Here's the structure of the dataset  
str(testdata)

## tibble [71 × 2] (S3: tbl\_df/tbl/data.frame)  
## $ test1 : num [1:71] 0.825 0.975 0.75 0.325 0.725 0.925 0.65 1 0.9 0.525 ...  
## $ perusall: num [1:71] 92.4 98.2 100 27.8 100 ...

#Here are some summary statistics on the variables  
summary(testdata)

## test1 perusall   
## Min. :0.3250 Min. : 0.00   
## 1st Qu.:0.6250 1st Qu.: 64.59   
## Median :0.7250 Median : 91.09   
## Mean :0.7222 Mean : 78.55   
## 3rd Qu.:0.8500 3rd Qu.: 99.14   
## Max. :1.0000 Max. :100.00

Here is the standard deviation:

sd(testdata$test1)

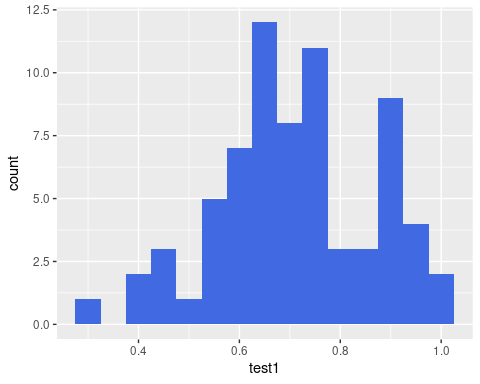
## [1] 0.1526872

sd(testdata$perusall)

## [1] 27.04033

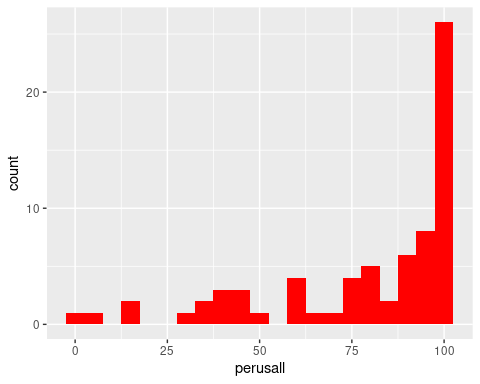
Here is a histogram for test1.

test\_hist <- ggplot(testdata, aes(x=test1))  
test\_hist + geom\_histogram(binwidth=.05, fill="royal blue")



Here is a histogram for perusall.

test\_perusall <- ggplot(testdata, aes(x=perusall))  
test\_perusall + geom\_histogram(binwidth=5, fill="red")



## 1. How well did students do on the test? Using all four pieces of R (the data structure, the summary statistics, the standard deviations, and the histogram), evaluate class performance. Be sure to frame your answer in terms of the 1) shape and 2) spread (dispersion) of the distribution and 3) measures of central tendency. (4 points)

## 2. Now look at the data for the variable perusall. What do you notice about the distribution, spread, and measures of central tendency for this variable? What does this information tell you about students’ reading habits and test scores in the class? (4 points)

## 3. For an instructor interested in making sure that test scores accurately reflect student effort, do you think that an average grade of 90 or higher on the readings is a good cutoff for evaluation? Why or why not? (2 points)

# Exercise 2: Hypothesis Generation (6 points)

I’m creating two groups to compare by separating students who did most of the reading (an average grade on readings on Perusall of at least 90 points (90%)) from those who did less. Later in the lab, we will be comparing the mean test grades for these two groups.

testdata <- mutate(testdata,  
 read=factor(ifelse(perusall>=90, 1, 0)))

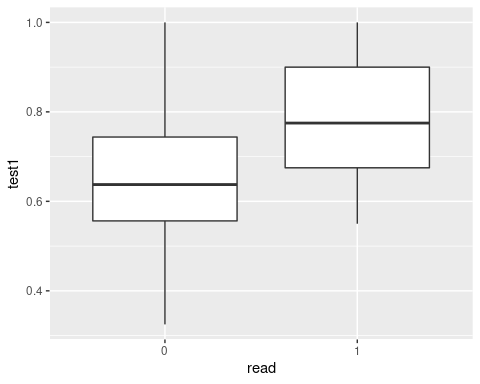
## 1. Write three hypotheses about the relationship between test scores (test1) and read with test scores as the dependent (outcome) variable:

### -a null hypothesis (2 points)

### -a two-tailed (non-directional) hypothesis (2 points)

### -a one-tailed hypothesis (2 points)

test\_read <- ggplot(testdata, aes(x=read, y=test1))  
test\_read + geom\_boxplot()

 ## 2. In a moment we’ll be comparing the mean test scores for the two groups, but the above boxplot shows us the distributions. Compare the boxplots for both groups. What do you notice about the relationship between test scores and doing 90% or more of the reading (read=1) and doing less than 90% of the reading (read=0)?

# Exercise 3: Reading R Output and Comparing the Means (10 points)

t.test(test1 ~ read, data=testdata, paired=FALSE, alternative="two.sided")

##   
## Welch Two Sample t-test  
##   
## data: test1 by read  
## t = -3.8888, df = 60.146, p-value = 0.0002542  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.19693700 -0.06315839  
## sample estimates:  
## mean in group 0 mean in group 1   
## 0.6544118 0.7844595

## 1. Interpret the output for the t-test comparing the means for students who did 90% or more of the reading compared to those who didn’t, reporting on the 95% confidence interval of the difference, the type of test (one-tailed or two-tailed), and what the results mean for the accepting or rejecting the null hypothesis and the alternative hypothesis. (8 points)

## 2. Based on these results, what would you advise students about the relationship between completing the reading assignments and performance on tests in this course? (2 points)