PSTAT 105 HW 3

Matthew Xu (5752811)

29 Janurary 2021

```r
library(rvest)
library(xml2)
library(stringi)

## set the working directory as the file location
setwd(getwd())

# read in data from file
shoulders <- read.table("shoulder.txt", header = TRUE)

earthquake <- read.table("EarthquakeData.htm", sep = ",", skip = 13, header = TRUE,
    nrows = 1826, colClass = c("integer", "integer", "integer", "character", "numeric",
        "numeric", "numeric", "numeric", "character"))
earthquakeHour <- as.integer(substr(earthquake$Time.hhmmss.mm.UTC, 1, 2))
earthquakeMin <- as.integer(substr(earthquake$Time.hhmmss.mm.UTC, 3, 4))
earthquakeSec <- as.numeric(substr(earthquake$Time.hhmmss.mm.UTC, 5, 8))

BBallBDays <- read.delim("BBallBDays.txt", header = TRUE, sep = "")
```

All of these questions require analyzing the data using R. Please prepare a typed report showing how you completed the analysis. Your report should include the R input and output as well as the plots and answers to all of the questions. Be sure to state your conclusions for each question in terms of the data being studied.

1. The Probability Department decrees that all courses are to be graded in such a way that the grades (as percentages) follow a beta distribution with alpha = 6 and beta = 2. In Prof. Smirnov's course, he gives the following grades

40% 71.7% 81.7% 83.3% 93.3% 97.5%

a. "PROBLEM NUMBER ONE, PART A" Draw a plot comparing the CDF of the beta distribution to the Empirical CDF of this data.

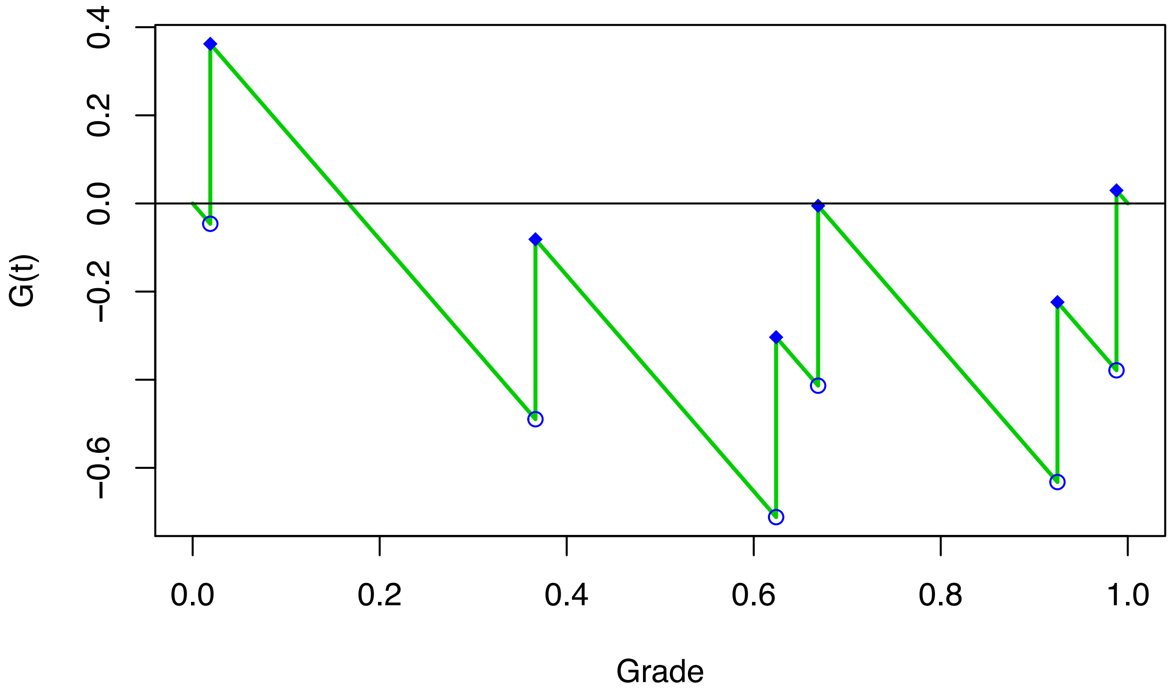
```
dev.off()
```

```
## null device
##           1
```

```
library(tidyverse)
plot.new()
percentages <- c(0.4, 0.717, 0.817, 0.833, 0.933, 0.975)

# empirical CDF
ecdf <- tibble(Observations = sort(percentages), CDF = seq(from = 1/6, to = 1, by = 1/6),
    Norm = sort(pnorm(percentages, mean = mean(percentages), sd = sd(percentages))))

ecdfPlot <- ggplot(data = ecdf, aes(x = Observations, y = CDF)) + geom_step(direction = "hv",
    col = "red") + labs(x = "Percentages", y = "Empirical CDF") + geom_line(mapping = aes(x =
    y = Norm), col = "green")
ecdfPlot

# plot cdf
points(percentages, pbeta(percentages, 6, 2), ylab = "density", type = "l", col = 4)
```

b. "PROBLEM NUMBER ONE, PART B" Draw a plot of the function Gn(t) = sqrt(n)(Fn(t) - t) where Fn is the empirical CDF of the data once it has been transformed so that it looks uniform.

```
# change to uniform
uniformPer <- pbeta(percentages, shape1 = 6, shape2 = 2)

# plot the function G(t)
plot(c(0, rep(uniformPer, each = 2), 1), sqrt(6) * (rep(seq(0, 1, by = 1/6), each = 2) -
    c(0, rep(uniformPer, each = 2), 1)), type = "l", lwd = 2, col = 3, xlab = "Grade",
    ylab = "G(t)")

# plot points with it
points(uniformPer, (seq(0, 5/6, by = 1/6) - uniformPer) * sqrt(6), col = 4)  #open circles
points(uniformPer, (seq(1/6, 1, by = 1/6) - uniformPer) * sqrt(6), pch = 18, col = 4)  #closed
abline(h = 0)
```



c. "PROBLEM NUMBER ONE, PART C" Use the ks.test function to test the null hypothesis that the data is from a beta(6, 2) distribution and calculate an exact P-value.

```r
ks.test(percentages, "pbeta", 6, 2)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  percentages
## D = 0.2906, p-value = 0.5955
## alternative hypothesis: two-sided
```

Ho: The data is from a beta(6,2) distribution Ha: The data is not from a beta(6,2) distribution

Exact p value of test of beta distribution (6,2) is 0.5955

conclusion below

d. "PROBLEM NUMBER ONE, PART D" What do you conclude?

Because the pvalue of 0.5955 is larger the the significance level of 0.05, the null hypothesis cannot be rejected. There is not sufficient evidence to state that the data is not from a beta(6,2) distribution and it is not statistically significant. It is possible to conclude to accept the null hypothesis that it is from a beta(6,2) distribution.

This is also consistent with the graphs depicted above that the distribution of the data is similar to the beta(6,2) distribution

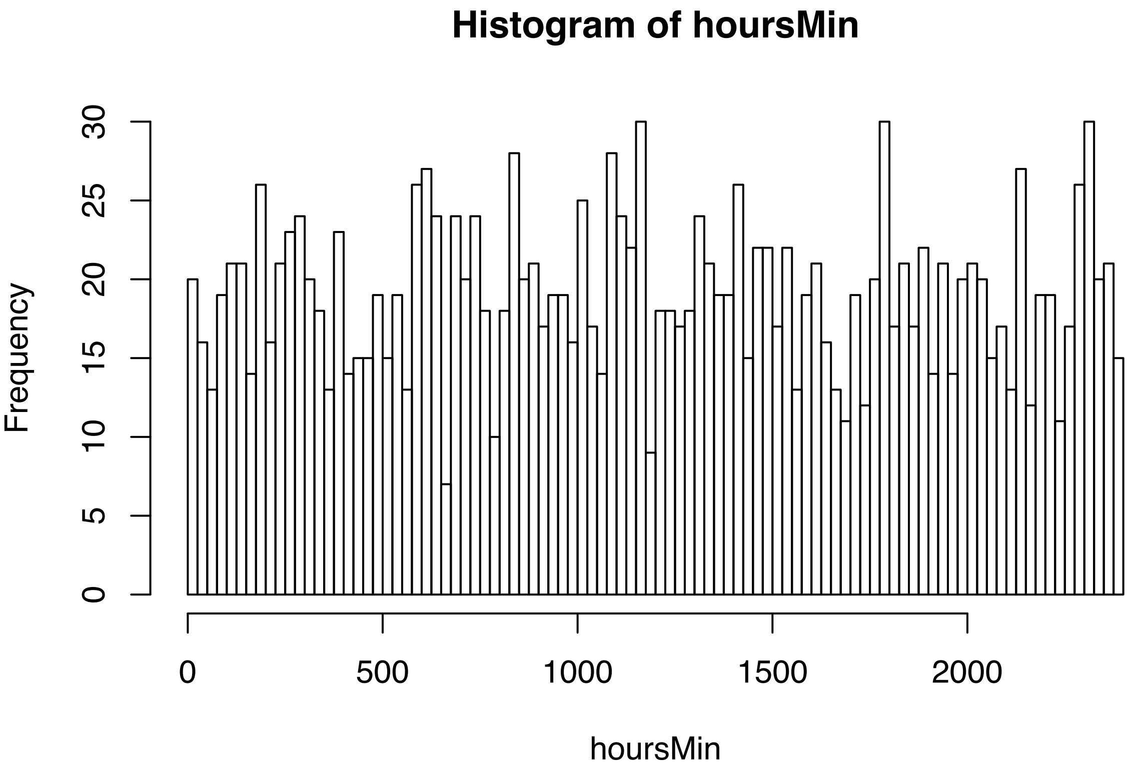
e. “PROBLEM NUMBER ONE, PART E” Why would it be difficult to apply a X2 test to this data?

The variable under study for the data is of percentages of grade scores, making it not categorical. In addition, although the sample size is greater than 5, it is still quite small at 6, making possible inconsistencies and irregularties in data.

2. I am interested in testing whether major earthquakes are more likely during certain times of the day. The USGS NEIC database provides a file EarthquakeData.htm which contains information about nearly all earthquakes of more than 6.0 on the Richter scale for dates from 2001 to 2012. You will need to do some data coding and interpretation to read in the times properly. I urge you to be careful in inputting the data.

a. “PROBLEM NUMBER 2, PART A” Plot a histogram of the times with bars 15 minutes wide.

```
hoursMin <- earthquakeHour * 100 + earthquakeMin/60 * 100

hist(hoursMin, breaks = seq(0, 2400, by = 15/60 * 100))
```



b. “PROBLEM NUMBER 2, PART B” Use a X2 test to determine if the hour of the day that the earthquakes originate is equally distributed among the 24 hours.

```
freqHour <- table(earthquakeHour)
percHour <- freqHour/1826
# if each hour is distributed euqally
expected <- 1/24

hourTS <- sum((percHour - expected)^2/expected)
hourTS
```

```
## [1] 0.009111423
```

```
# df = 24-1
pchisq(hourTS, df = 23, lower.tail = FALSE)
```

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

H0: Hours of the day that the earthquakes originate is equally distributed among 24 hours HA: Hours of the day that the earthquakes originate is not equally distributed among 24 hours

Because the p value is larger than the significance level of 0.05 and the test stastic is smaller than the critical value, we cannot reject the null hypothesis. There is not enough evidence to state that the hours of the day that the earthquakes originate is not equally distributed among 24 hours and that it is not stastically significant. It is possible to conclude to accept the null hypothesis that earthquake hours are evenyl distributed among 24 hours.

c. "PROBLEM NUMBER 2, PART C" Calculate the Kolmogorov–Smirnov statistics D+ and D- for testing that the times are uniformly distributed throughout the day.

```
time.frac <- earthquakeHour/24 + earthquakeMin/(24 * 60) + earthquakeSec/(24 * 60 *
    60)
up <- seq(1/1826, 1, by = 1/1826) - sort(time.frac)
max(up)
```

```
## [1] 0.01338519
```

```
min(up)
```

```
## [1] -0.01209321
```

d. "PROBLEM NUMBER 2, PART D" Calculate an approximate P–value for the appropriate test.

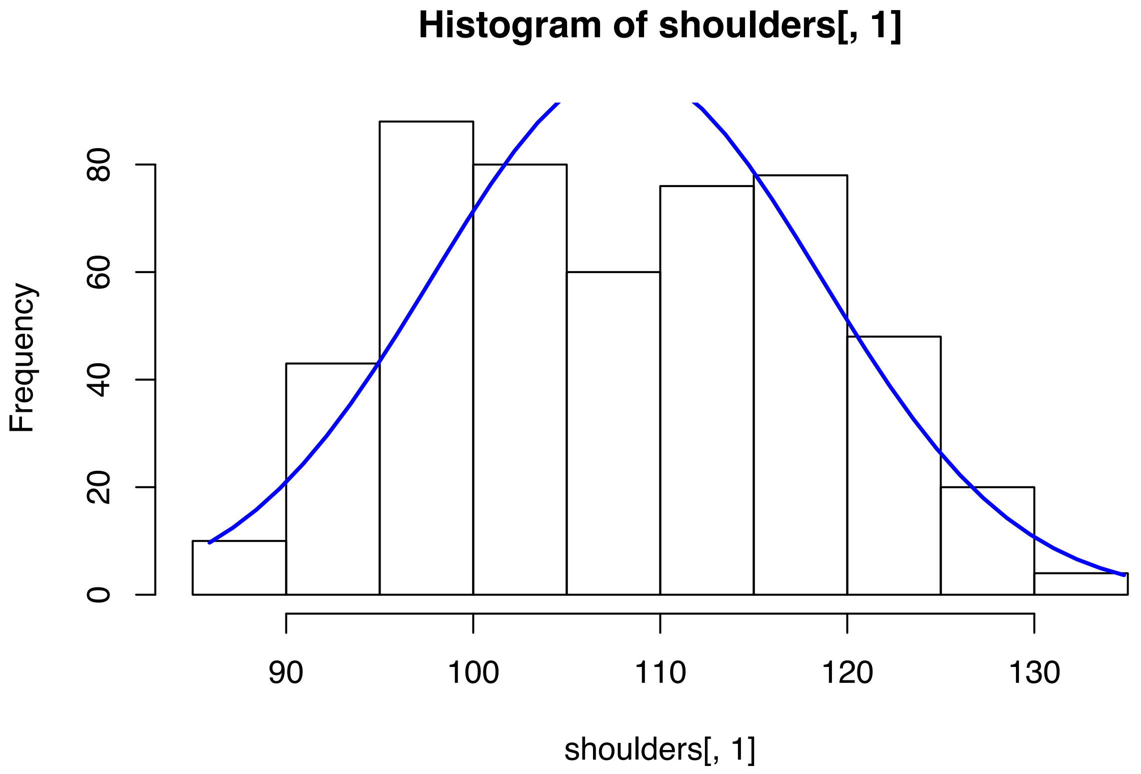
```
ks.test(as.integer(earthquake$Time.hhmmss.mm.UTC), "punif", 0, 24)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  as.integer(earthquake$Time.hhmmss.mm.UTC)
## D = 0.99945, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

3. Grete Heinz and Louis J. Peterson, at San Jose State University and at the U.S. Naval Postgraduate School in Monterey, California, took measurements from 507 subjects. Part of their data set is in the file shoulder.txt which includes the width of each subjects shoulders. We want to test if this data is normally distributed. We will be using the library nortest to perform these tests.

a. “PROBLEM NUMBER 3, PART A” Plot a histogram of the shoulder data using a number of breaks that you think is appropriate. Draw the density of a normal distribution with the same mean and variance over the histogram for comparison.

```
h <- hist(shoulders[, 1], breaks = 10)
xfit <- seq(min(shoulders[, 1]), max(shoulders[, 1]), length = 40)
yfit <- dnorm(xfit, mean = mean(shoulders[, 1]), sd = sd(shoulders[, 1]))
yfit <- yfit * diff(h$mids[1:2]) * length(shoulders[, 1])

lines(xfit, yfit, col = "blue", lwd = 2)
```



b. “PROBLEM NUMBER 3, PART B” Use lillie.test, cvm.test, and ad.test to test whether or not this data is from a normal distribution. What do you conclude?

```r
library(nortest)
lillie.test(shoulders[, 1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  shoulders[, 1]
## D = 0.077918, p-value = 8.838e-08
```

```r
cvm.test(shoulders[, 1])
```

```
##
##  Cramer-von Mises normality test
##
## data:  shoulders[, 1]
## W = 0.62429, p-value = 2.321e-07
```

```r
ad.test(shoulders[, 1])
```

```
##
##  Anderson-Darling normality test
##
## data:  shoulders[, 1]
## A = 3.6469, p-value = 4.112e-09
```

H0: The distribution of shoulder data is normally distributed HA: The distribution of shoulder data is not normally distributed

In the Lilliefors normality test, the p value is less then the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of shoulder data is not normally distributed and that it is statistically significant.

In the Cramer-von Mises normality test, the p value is less then the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of shoulder data is not normally distributed and that it is statistically significant.

In the Anderson-Darling normality test, the p value is less then the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of shoulder data is not normally distributed and that it is statistically significant.

c. “PROBLEM NUMBER 3, PART C” I’m concerned that the fact that men and women have different general body sizes may be causing a problem with our test. Re-run the normality tests on just the men and just the women separately. What do you conclude?

```r
men <- subset(shoulders, Gender == "Male", select = Shoulder:Gender)
women <- subset(shoulders, Gender == "Female", select = Shoulder:Gender)

lillie.test(men[, 1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  men[, 1]
## D = 0.041117, p-value = 0.3921
```

```r
cvm.test(men[, 1])
```

```
##
##  Cramer-von Mises normality test
##
## data:  men[, 1]
## W = 0.035563, p-value = 0.7605
```

```r
ad.test(men[, 1])
```

```
##
##  Anderson-Darling normality test
##
## data:  men[, 1]
## A = 0.2263, p-value = 0.8158
```

```r
lillie.test(women[, 1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  women[, 1]
## D = 0.076669, p-value = 0.0008286
```

```r
cvm.test(women[, 1])
```

```
##
##  Cramer-von Mises normality test
##
## data:  women[, 1]
## W = 0.28647, p-value = 0.0004713
```

```r
ad.test(women[, 1])
```

```
##
##  Anderson-Darling normality test
##
## data:  women[, 1]
## A = 1.6058, p-value = 0.0003889
```

H0: The distribution of men shoulder data is normally distributed HA: The distribution of men shoulder data is not normally distributed

In the Lilliefors normality test for men, the p value is greater than the significance level of 0.05, therefore we cannot reject the null hypothesis. There is not sufficient evidence to state that the distribution of shoulder data of men is not normally distributed and that it is not statistically significant. It is possible to conclude that the null hypothesis is accepted and that the distribution of men data is normally distributed.

In the Cramer-von Mises normality test for men, the p value is greater than the significance level of 0.05, therefore we cannot reject the null hypothesis. There is not sufficient evidence to state that the distribution of shoulder data of men is not normally distributed and that it is not statistically significant. It is possible to conclude that the null hypothesis is accepted and that the distribution of men data is normally distributed.

In the Anderson-Darling normality test for men, the p value is greater than the significance level of 0.05, therefore we cannot reject the null hypothesis. There is not sufficient evidence to state that the distribution of shoulder data of men is not normally distributed and that it is not statistically significant. It is possible to conclude that the null hypothesis is accepted and that the distribution of men data is normally distributed.

H0: The distribution of women shoulder data is normally distributed HA: The distribution of women shoulder data is not normally distributed

In the Lilliefors normality test for women, the p value is less than the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of shoulder data of women is not normally distributed and that it is statistically significant.

In the Cramer-von Mises normality test for women, the p value is less than the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of shoulder data of women is not normally distributed and that it is statistically significant.

In the Anderson-Darling normality test for women, the p value is less than the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of shoulder data of women is not normally distributed and that it is statistically significant.

It appears that seperatly, the two sets of shoulder data follow different distributions. However, together they appear to not be normally distributed, indicating that perhaps the women data has larger weight and influence when calculting test stastices for these normality tests, possibly due to a larger sample (260 women 247 men). The difference between the data in mens and data in womens is different enough of a trend that they are not similar.

d. "PROBLEM NUMBER 3, PART D" I hate to divide a data set into two parts. I really want to perform one test using all of the data. We could calculate the mean for the male subjects and the mean for the female subjects. Then, subtract these means from the data points so that we move the two populations on top of each other. Try the normality tests on this new set of data. What do you conclude?

```
men_mean = mean(men[, 1])
women_mean = mean(women[, 1])

for (i in 1:507) {
    if (shoulders[i, 2] == "Male") {
        shoulders[i, 3] = shoulders[i, 1] - men_mean
    } else {
        shoulders[i, 3] = shoulders[i, 1] - women_mean
    }
}

lillie.test(shoulders[, 3])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  shoulders[, 3]
## D = 0.041835, p-value = 0.03391
```

```
cvm.test(shoulders[, 3])
```

```
##
##  Cramer-von Mises normality test
##
## data:  shoulders[, 3]
## W = 0.16851, p-value = 0.01357
```

```
ad.test(shoulders[, 3])
```

```
##
##  Anderson-Darling normality test
##
## data:  shoulders[, 3]
## A = 0.94516, p-value = 0.01668
```

H0: The difference between the shoulder data and its respective gender mean is normally distributed
HA: The difference between the shoulder data and its respective gender mean is not normally distributed

In the Lilliefors normality test, the p value is less then the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of difference in shoulder data and the gender means is not normally distributed and that it is statistically significant.

In the Cramer-von Mises normality test, the p value is less then the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of difference in shoulder data and the gender means is not normally distributed and that it is statistically significant.

In the Anderson-Darling normality test, the p value is less then the significance level of 0.05, therefore we reject the null hypothesis. There is sufficient evidence to state that the distribution of difference in shoulder data and the gender means is not normally distributed and that it is statistically significant.

By subtracting the respective gender means from the data, the combined data set of differences of both genders makes the same inital conclusion that the data is not normally distributed. It is possible that this method has little influence on the data points, as the tranformation is linear and on its own gender subset of the data, not considering the other gender. Therefore, womens data might still have higher influence possibly due to its larger sample size than mens (women:260, men:247).

e. "PROBLEM NUMBER 3, PART E" Do you think it would be reasonable to assume a normal distribution and perform a two-sample t-test using this data? Why?

```r
var(men[, 1])
```

```
## [1] 42.22431
```

```r
var(women[, 1])
```

```
## [1] 41.86863
```

Due to the Central Limit thereom and the large number of samples (ex: >= 30), the distribution of the data can be generalized to a normal distribution. The data appears to be continous (shoulder sizes), follows a normal distribution by the CLT, the variances of both population men and women appear to be similar, abd the two samples appear to be inpdependent and assumed random sampleled. Therefore, a two-sample t-test may be reasonable.

4. Use a Kolomogorov–Smirnov test to test whether the distribution of birthdays in the basketball data from last week is uniformly distributed across the year. Please describe your analysis and conclusion.

```r
library(date)
BBallBDays$x <- paste(BBallBDays$Month, BBallBDays$Day, BBallBDays$Year)
# BBallBDays$Dates <- as.date(BBallBDays$x,'%m/%d/%Y') ks.test(BBallBDays,
# 'punif',mean=sapply(BBallBDays, mean), sd=sapply(BBallBDays, sd))
```