PSTAT 105 HW 2 Question 2 and 3

Matthew Xu (5752811)

22 Janurary 2021

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

# read in data from file
Selltimes <- scan("Selltimes.txt")
```

For Questions 2 and 3, please analyze the data using R and type up your answers to these questions.

2. A well-known analysis in Malcolm Gladwell's book Outlier argues that the best hockey players are more likely to be born earlier in the year presumably because this gives them advantages in the youth hockey leagues. We are interested in checking whether there is a similar effect in basketball.

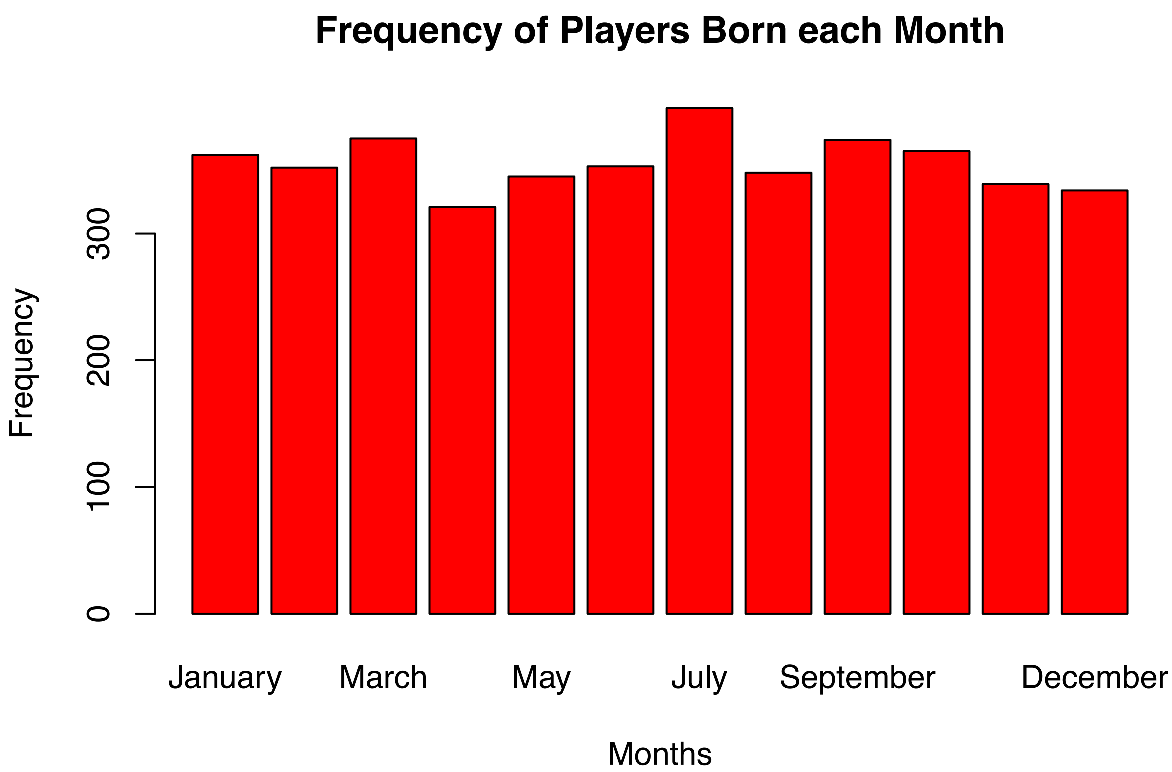
a. "PROBLEM NUMBER TWO, PART A" The data set BballDays.txt contains the names and date of birth for a large sample of professional basketball players listed on the http://www.basketball-reference.com web site. Use the table function to calculate how many players were born in each month. Draw an appropriate plot

```r
# read in data from file
BBallBDays <- read.delim("BBallBDays.txt", header = TRUE, sep = "")

# Use the table function to calculate how many players were born in each month.
monthFreq <- table(BBallBDays$Month)
monthFreq <- monthFreq[month.name]
monthFreq
```

```
##
##   January  February     March     April       May      June      July    August
##       362       352       375       321       345       353       399       348
## September   October  November  December
##       374       365       339       334
```

```r
# Draw an appropriate plot
barplot(monthFreq, main = "Frequency of Players Born each Month", xlab = "Months",
    ylab = "Frequency", col = "red")
```



b. "PROBLEM NUMBER TWO, PART B" Perform a X^2 test to see if the players are equally likely to be born in any month

```r
numPlayers <- sum(monthFreq)
expected <- numPlayers/12
testStat <- sum((monthFreq - expected)^2/expected)
testStat
```

```
## [1] 13.65901
```

```r
# critical value calculation df = n-1 = 12-1 = 11
qchisq(0.95, df = 11)
```

```
## [1] 19.67514
```

```r
pchisq(testStat, df = 11)
```

```
## [1] 0.7475608
```

Assuming each player is equally likely to occur in each month, of the total 4267 total players (players with NA on their birthdays are removed), the average number of players born each month is 4267/12 = 358.25, which is our expected value for the chi-square test statistic. The observed number of players per month is in the table monthFreq. Degrees of freedom = n-1, 12-1 = 11.

H0 = Players are equally likely to be born each month, HA = Players are not equally likely to be born each month

Because the test statistic of 13.65901 is less than the critcal value of 19.67514 and the pvalue is larger than significance level of 0.05, we do not reject the null hypothesis. There is not enough evidence to indicate that the number of players born each month is different and not equally likely and that it is not statically significant. It is possible to accept the null hypothesis and assume the possiblity that each month can have the same numebr of players being born.

c. "PROBLEM NUMBER TWO, PART C" In order to focus our attention on modern players, repeat this analysis with only those players that were born after 1/1/1955. (also use this smaller data set for the following questions.)

```
after1955 <- BBallBDays[BBallBDays$Year >= 1955, ]
after1955_table <- table(after1955$Month)
after1955_table <- after1955_table[month.name]
after1955_table
```

```
##
##   January  February     March     April       May      June      July    August
##       208       226       222       197       220       219       217       201
## September   October  November  December
##       220       225       197       195
```

```
after1955_numPlay <- sum(after1955_table)
after1955_expected <- after1955_numPlay/12
after1995_testStat <- sum((after1955_table - after1955_expected)^2/after1955_expected)
after1995_testStat
```

```
## [1] 7.266196
```

```
pchisq(after1995_testStat, df = 11, lower.tail = FALSE)
```

```
## [1] 0.7771349
```

For players born after 1955:

H0 = Players are equally likely to be born each month for players born after 1/1/1995, HA = Players are not equally likely to be born each month for players born after 1/1/1995

Because the test statistic of 7.266196 is less than the critcal value of 19.67514 and the p values is larger than alpha of 0.05, we do not reject the null hypothesis. There is not enough evidence to indicate that the number of players born each month is different using players born after 1/1/1955 and that it is not statistically significant. It is possible to accept the null hypothesis and assume the possiblity that each month can have the same numebr of players being born.

d. "PROBLEM NUMBER TWO, PART D" To be more careful, we should realize that more people are probably born in January than Feburary just because there are more days in January. Perform a X^2 test where the null hypothesis is that the probability of each month is proportional to the average number of days in that month.

```r
# find total of the subset after1955 data
n_after <- sum(after1955_table)
# 28.25 is for Feburary every 4 years average is a leap year
p_after <- c(31, 28.25, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)/365.25
e_after <- n_after * p_after
x_after <- sum((after1955_table - e_after)^2/e_after)
x_after
```

```
## [1] 10.74596
```

```r
pchisq(x_after, df = 11, lower.tail = FALSE)
```

```
## [1] 0.4647846
```

Ho: For players after 1/1/1995 birthdays probaibility of each month is proportional to average number of days in that month Ha: For players after 1/1/1995 birthdays probaibility of each month is not proportional to average number of days in that month

Because the P-value of 0.4647846 is larger than alpha of 0.05 and the test statstic is smaller than the critcal value for Chi-square at 0.05 significance level and 11 df of 19.675, we do not reject the null hypothesis. There is not sufficient evidence to state that each month is not proportional to average number of days in that month for players born after 1/1/1995. It is statiscally not significant. It is possible to accept the null hypothesis.

e. "PROBLEM NUMBER TWO, PART E" Going even further, it seems that some months generally are favored over others for having babies (summer births are more likely). We should probably compare our basketball player data to the following probabilities from the CDC.

Month Jan Feb Mar Apr May Jun Prob. 0.0815 0.0752 0.0837 0.0816 0.0859 0.0813 Month Jul Aug Sep Oct Nov Dec Prob. 0.0883 0.0892 0.0866 0.0849 0.0787 0.0830

Perform a X^2 test to see if the basketball player data has the same distribution.

```r
p_cdc <- c(0.0815, 0.0752, 0.0837, 0.0816, 0.0859, 0.0813, 0.0883, 0.0892, 0.0866,
    0.0849, 0.0787, 0.083)
e_cdc <- n_after * p_cdc
ts_cdc <- sum((after1955_table - e_cdc)^2/e_cdc)
ts_cdc
```

```
## [1] 12.81138
```

```r
pchisq(ts_cdc, df = 11, lower.tail = FALSE)
```

```
## [1] 0.3058319
```

f. "PROBLEM NUMBER TWO, PART F" Interpret your results. Is there significant evidence at an a = 0.05 level that professional basketball players are born earlier in the year than the normal population?

Ho: player monthly birth data follows the CDC data and distribution Ha: player monthly birth data does not follow the CDC data and distribution

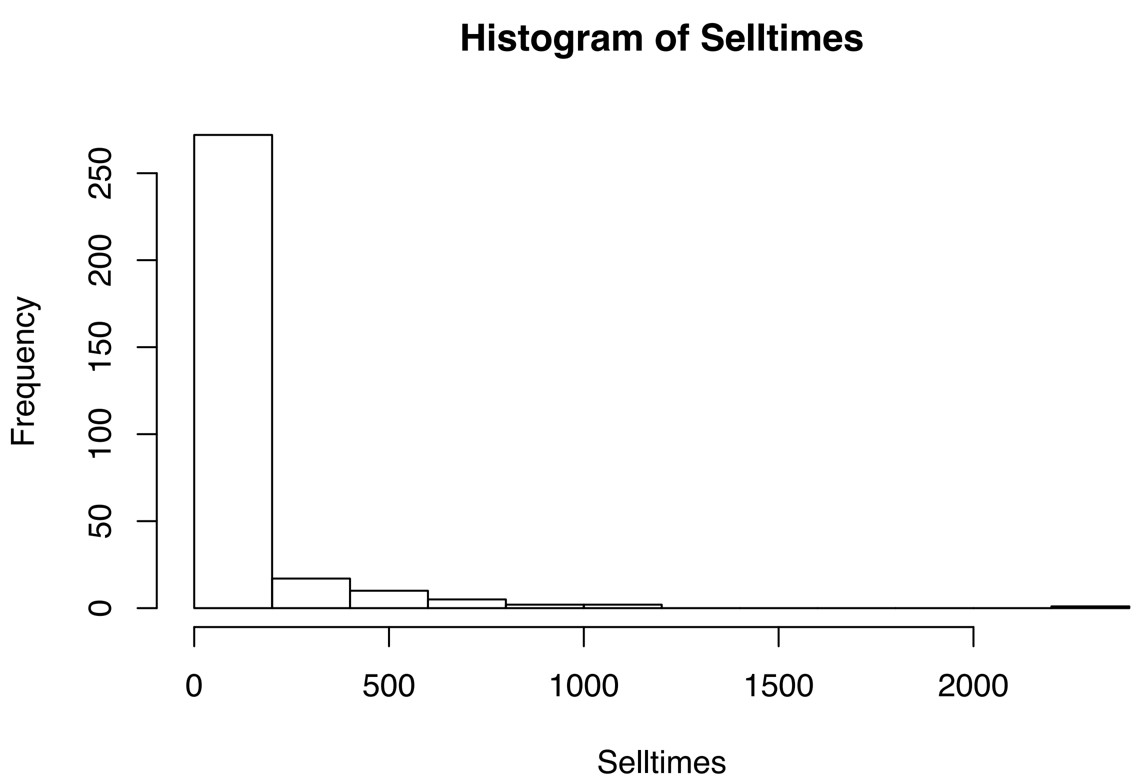
Because the p value of 0.3058319 is larger than the significance level of 0.05 and test statistic is less than the critical value at signifcance level 0.05 and df = 11, we do not reject the null hypothesis. There is not sufficient evidence to state that player monthly birth data does not follow CDC data and distribution and that it is not stastically significant. It is possible to accept the null hypothesis.

3. The data set Selltimes.txt consists of the time that elapses between when sell orders for CISCO stock

were placed during April 5, 2010. My hypothesis is that these times have an exponential distribution with CDF: F(t) = 1 - e^-lambda*t for some unknown rate lambda.

a. Use the hist function to plot an informative histogram of the data.

```r
hist(Selltimes)
```



b. Calculate the MLE, lambda = x^-1, from the data.

```
# find MLE
xBar = mean(Selltimes)
MLE = xBar^-1
MLE
```

```
## [1] 0.01320614
```

c. Use this estimate of lambda to divide the sample space into 10 intervals that will be big enough that the X^2 approximation will be appropriate.

```
# 2400 is about the max data point 0-200 is about the max amount for each of the
# 10 bins
n <- length(Selltimes)
intervals <- c(seq(0, 200, length = 10), 2400)
cdf <- pexp(intervals[c(-1, -11)], MLE)
exp <- (c(cdf, 1) - c(0, cdf)) * n
exp
```

```
##  [1] 78.587441 58.600432 43.696684 32.583381 24.296506 18.117217 13.509496
##  [8] 10.073649  7.511635 22.023558
```

d. Use the hist function to count the number of observations in each of those intervals

```
# use attribute counts from list
sell.counts <- hist(Selltimes, breaks = intervals, plot = F)
sell.counts$counts
```

```
##  [1] 226  14   6  10   5   4   2   3   2  37
```

e. Perform the appropriate X^2 test

```
# df = k-p-1 = 10-1-1 = 8
sell_ts <- sum((sell.counts$counts - exp)^2/exp)
pchisq(sell_ts, 8, lower.tail = FALSE)
```

```
## [1] 1.932279e-84
```

H0: Time from Selltimes follow an exponenetial distribution for 10 intervals Ha: Time from Selltimes does not follow an exponential distribution for 10 intervals

Because the P-value of 1.932279e^-84 is less than alpha of 0.05, we reject the null hypothesis. There is sufficient evidence to state that time from Selltimes does not follow an exponential distribution and that it is statistically significant.

f. Inspect the counts and the expected values and give some description of how the data looks different from an exponential distribution.

```r
print((sell.counts$counts - exp)^2/exp)
```

```
##  [1] 276.513169  33.945117  32.520545  15.652431  15.325460  11.000355
##  [7]   9.805584   4.967069   4.044142  10.184268
```

```r
print(sell.counts$counts)
```

```
##  [1] 226  14   6  10   5   4   2   3   2  37
```

```r
print(exp)
```

```
##  [1] 78.587441 58.600432 43.696684 32.583381 24.296506 18.117217 13.509496
##  [8] 10.073649  7.511635 22.023558
```

As you can see, the sell.counts$counts derivied from the interval does not follow the estimated exponential distribution of the data, therefore making it different from the exponential distribution. In addition, the chi-square test shown in the previous question also indicates this, as the null hypothesis is rejected and concludes that the data is not exponetial.

g. What difference does it make if we used 25 or 100 intervals instead of 10? Experiment a little with different sets of intervals and report the results and whether they demonstrate anything different from the original 10-interval analysis

```r
# 25 intervals
intervals_25 <- c(seq(0, 200, length = 25), 2400)
cdf_25 <- pexp(intervals[c(-1, -11)], MLE)
exp_25 <- (c(cdf_25, 1) - c(0, cdf_25)) * n

sell.counts25 <- hist(Selltimes, breaks = intervals_25, plot = F)
sell.counts25$counts
```

```
##  [1] 201  16  11   5   4   7   1   1   4   5   1   2   3   1   3   0   0   1   2
## [20]   1   1   1   1   0  37
```

```r
sell_ts25 <- sum((sell.counts25$counts - exp_25)^2/exp_25)
# df = 25-1-1 = 23
pchisq(sell_ts25, 23, lower.tail = FALSE)
```

```
## [1] 2.478211e-159
```

```r
# 100 intervals
intervals_100 <- c(seq(0, 200, length = 100), 2400)
cdf_100 <- pexp(intervals_100[c(-1, -11)], MLE)
exp_100 <- (c(cdf_100, 1) - c(0, cdf_100)) * n

sell.counts100 <- hist(Selltimes, breaks = intervals_100, plot = F)
sell.counts100$counts
```

```
##   [1] 163  22   8   6   7   2   2   5   4   2   5   1   1   2   2   0   1   1
##  [19]   1   0   3   2   0   3   1   0   0   1   0   0   0   0   1   1   0   2
##  [37]   1   2   2   0   1   0   1   0   0   1   0   1   0   1   1   1   0   0
##  [55]   0   0   1   0   1   2   0   0   0   0   0   0   0   0   0   0   0   0
##  [73]   1   0   0   1   0   1   0   1   0   0   0   0   0   0   1   0   0   1
##  [91]   0   0   1   0   0   0   0   0   0  37
```

```r
sell_ts100 <- sum((sell.counts100$counts - exp_100)^2/exp_100)
# df = 100-1-1 = 98
pchisq(sell_ts100, 98, lower.tail = FALSE)
```

```
## [1] 0
```

Ho: For 25 or 100 intervals, the data is exponetially distributed Ha: For 25 or 100 intervals, the data is not exponetially distributed

For both intervals of 25 and 100, the p values are much smaller than the significance level of 0.05, we reject the null hypothesis. There is sufficient evidence to state that for both 25 and 100 intervals, the distribution is not exponential and that it is statistically significant.

In fact, as the number of intervals increases, the p value continues to become smaller and smaller, making it even more evident that when intervals increase, it cannot be generalized by a exponenetial distribution