STAT5703 HW2 Ex3

Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

Exercise 3

Q1

The estimated values of the transition matrix are given by the observations, so for example, $\hat{p_{00}}$ is the number of two consecutive snowy days divided by the total number of snowy day. So in Charleston, the probability of whether the weather next day is the same with that of today is similar no matter what day today is. For Medford, it's more likely to see a snowy day after a no snow day than the opposite. While for New York, it's more likely to see a no snow day after a snowy day than the opposite.

Q2

The probability would be $\hat{p_{11}}$ in the $\mathbf{T}_{NY}$ matrix, which is 0.776.

Q3

The long term probability is the stationary distribution of the Markov chain. So solving the equation $\pi P = P$, we have $\pi = (\pi_0, \pi_1) = (0.0733, 0.9267)$. So the long term probability of having a snow day in Medford is 0.9267.

Q4

Since the these two conditional distributions are binomial distributions, we can denote their success rate as p_1 and p_2 respectively. Here success means the random variable in front of the condition takes the value of 1. Then We can formulate the null hypothesis H_0 as, these two distributions are not significantly different, i.e., $p_1 = p_2$. So by using the approximation of normal distribution, we can use a z-statistic to test this hypothesis. Under H_0, the test statistic will have the distribution as,

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - 0}{\sqrt{\hat{p}(1-\hat{p})\,(1/n_1 + 1/n_2)}} \sim N(0,1)$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2}$. Given that number of snowy days is 1000 and total sample size is 1373, we have,

$$p = \frac{1000 * 0.955 + 373 * 0.966}{1373} = 0.9580$$

Then the exact value for z-statistic will be,

$$\frac{|0.955 - 0.966|}{\sqrt{0.9580 * (1 - 0.9580) * (1/1000 + 1/373)}} = 0.9038$$

and therefore the p-value is 0.3661, which means that we can not reject the null hypothesis that these two distributions are not significantly different.

Q5

```
library(markovchain)
library(lubridate)
library(dplyr)
```

```
parse_year <- function(x, start_year=1911) {
  # assign right century for the year
  x <- mdy(x)
  y <- year(x) %% 100
  year(x) <- ifelse(y > start_year %% 100, 1900+y, 2000+y)
  x
}

centralPark <- read.csv("./CentralPark.csv")

generate_chains <- function(row_range, prefix=1900) {
  snow <- centralPark[row_range, ] %>%
    mutate(DATE_OBJ=parse_year(DATE, start_year = prefix)) %>%
    arrange(DATE_OBJ) %>%
    filter(!is.na(SNWD)) %>%
    mutate(SNOW_DAY=factor(ifelse(SNWD < 50, "NoSnow", "SnowDay"))) %>%
    mutate(YEAR=year(DATE_OBJ)) %>%
    filter((month(DATE_OBJ) == 12 &
                        day(DATE_OBJ) >=17 &
                        day(DATE_OBJ) <= 31)) %>%
    group_by(YEAR) %>%
    summarise(Weather=list(SNOW_DAY), Count=n()) %>%
    filter(Count == 15)
  snow
}



# after row 36524, we are in 21st century
snow19 <- generate_chains(1:36524)
snow20 <- generate_chains(36525:nrow(centralPark))

snow <- rbind(snow19, snow20)
```

```
mcFit <- markovchainFit(snow$Weather, "mle")
mcFit$estimate
```

```
## MLE Fit
##  A  2 - dimensional discrete Markov Chain defined by the following states:
##  NoSnow, SnowDay
##  The transition matrix  (by rows)  is defined as follows:
##            NoSnow    SnowDay
## NoSnow  0.9698697 0.03013029
## SnowDay 0.2215190 0.77848101
```

Here we use as much data as we can, and filter out those years with inadequate records. Since we have 99*15=1485 sample points in total, which is larger than that used by the original estimate, the results may be slightly different. As shown by the result, every entry in transition matrix is as accurate up to two digits behind the dot, so the reproduction is successful.

Q6

I don't think a higher order chain help improve the results. We can justify it using the AIC value or likelihood ratio test. Here we choose the AIC value as the metric. However, the SMPracticals package can not calculate the likelihood across several realizations of a single markov chain. To accurately calculate the likelihood function across several relizations, we should first count the number of occurance of each pattern, which can be quite tedious when dealing with high order models. So here I simply merge all realizations into one long sequence. Though it will introduce bias when dealing with data points at the start or end of original realizations, it will reflect the pattern of the majority of data points as well.

```
library(SMPracticals)

merge_seq <- unlist(cbind(snow$Weather))

MClik(merge_seq)$AIC
```

```
## [1] 1057.9265  607.1833  610.8622  615.5379
```

So the AIC score shows that the first order model has the lowest AIC value and higher order models don't improve the fit.