Math 490 Exam #3

Maxwell Levin

April 14, 2018

Problem 1 (15 pts)

Most vehicles have devices measuring various quantities related to performance. One of these is fuel efficiency, or gas milage, usually expressed as miles per gallon (mpg). For one vehicle, the mpg values were recorded each time the gas tank was filled, and the mpg measuring device was then reset. Here are the mpg values for a random sample of 20 of these records.

```
 [1] 41.5 50.7 36.6 37.3 34.2 45.0 48.0 43.2 47.7 42.2 43.2 44.6 48.4 46.4
[15] 46.8 39.2 37.3 43.5 44.3 43.3
```

a. Construct a 95% Student's t-confidence interval for the true average mpg of the vehicle. Report the standard error used in this t-confidence interval.

We can get a 95% confidence interval for the mean mpg in R by simply asking R to run a t-test on our mpg data:

```
t.test(mpg)
```

```

    One Sample t-test

data:  mpg
t = 43.729, df = 19, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 41.10375 45.23625
sample estimates:
mean of x
    43.17
```

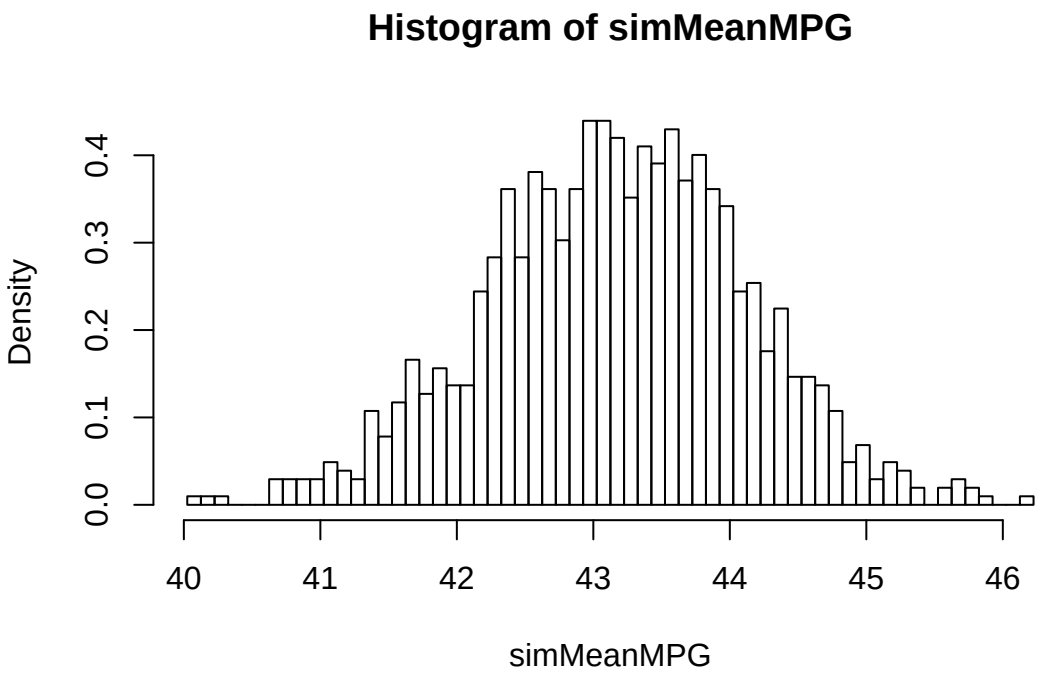
Thus we see that a 95% confidence interval for the true mean mpg of the vehicle is the interval from 41.10 to 45.24. Additionally, we note that the mean mpg from our sample data is 43.17.

b. Bootstrap the sample mean, $\overline{X}$, using $B = 1024$ resamples from the original mpg values. Make a histogram of the bootstrap sampling distribution. Also describe the shape and center of the distribution, and report the bootstrap standard error.

We can run the following code in R to bootstrap the mean of mpg values:

```
bootMean = function(data, rep) {
  replicate(rep, mean(sample(data, length(data), replace=TRUE)))
}

simMeanMPG = bootMean(mpg, 1024)
hist(simMeanMPG, breaks=seq(min(simMeanMPG), max(simMeanMPG) + 0.1, 0.1), probability=TRUE)
```



The bootstrap standard error is:

```
[1] 0.9512431
```

The mean of our boostrap distribution is

```
[1] 43.17375
```

We see that our boostraped mean mpg distribution is roughly normal with a mean of about 43.2 (about the same as our sample distribution) and that it has a standard error of about 0.9.

c. Use the bootstrap samples in part (b) to construct a 95% bootstrap t-confidence interval for the true average mpg of the vehicle. Would it be appropriate to report such a bootstrap t-confidence interval? Explain your answer.

We can run the following code in R to get a 95% confidence interval for the average mpg of the vehicle:

```
lower = mean(simMeanMPG) - abs(qt(.975, length(simMeanMPG) - 1) * sd(simMeanMPG))
upper = mean(simMeanMPG) + abs(qt(.975, length(simMeanMPG) - 1) * sd(simMeanMPG))
c(lower, upper)
```

```
[1] 41.30714 45.04037
```

We can test this confidence interval by running the following code in R:

```
length( simMeanMPG[ simMeanMPG >= lower & simMeanMPG <= upper]) / length(simMeanMPG)
```

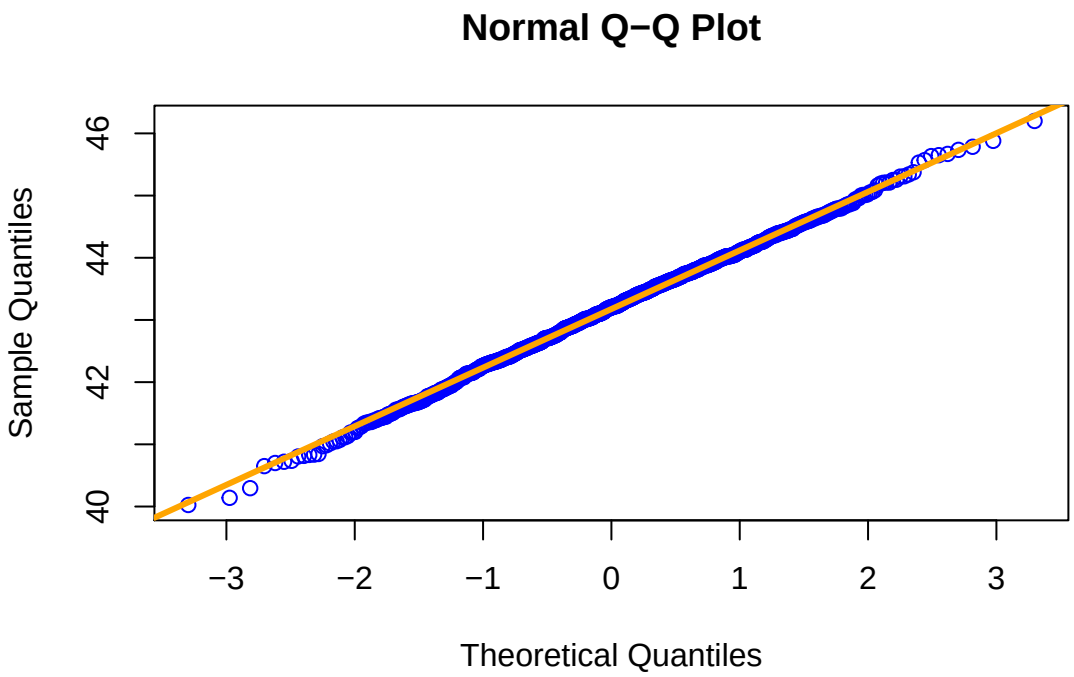
```
[1] 0.9511719
```

Thus we see that we have accurately created a 95% confidence interval for the bootstrapped standard deviation.

Our bootstrap distribution, and hence our confidence interval, will be reasonably representative of our original sample if our bootstrap distribution is normally distributed. We have several options for how to check

normality. I will proceed by making a normal quantile plot because I think it looks nice. We run the following code in R:

```
qqnorm(simMeanMPG, col="blue")
qqline(simMeanMPG, col="orange", lwd=3)
```



We see that the data is close to the normal quantile line. This indicates that our data could be normally distributed. To double check this I think it is appropriate to run a Shapiro-Wilk test:

```
    Shapiro-Wilk normality test

data:  simMeanMPG
W = 0.99883, p-value = 0.7548
```

The large P-value in our Shapiro-Wilk test tells us that our bootstrapped mean mpg distribution is normally distributed. Because of this I would say that it is appropriate to report our 95% confidence interval for our bootstrapped mean mpg, but that we should still be cautious of such a statistic due to the low sample size for our mpg.

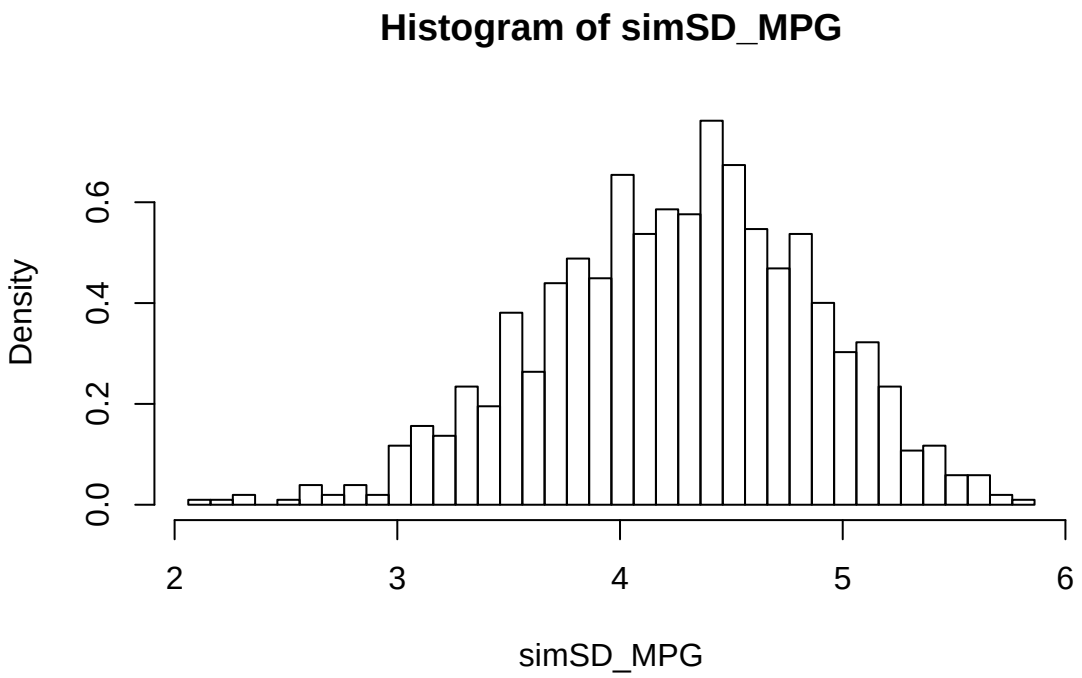
In addition to the average mpg, the driver is also interested in how much variablility (measured by standard deviation) there is in the mpg. The sample standard deviation S would vary from sample to sample, but we have no formula for the standard error of S.

d. Bootstrap the sample standard deviation by using 1024 resamples from the original mpg values. Make a histogram of the bootstrap sampling distribution. Be sure to describe the shape and the center of the distribution, and report the bootstrap standard error.

We can run the following code in R to bootstrap the sample standard deviations:

```r
bootSD = function(data, rep) {
  replicate(rep, sd(sample(data, length(data), replace=TRUE)))
}

simSD_MPG = bootSD(mpg, 1024)
hist(simSD_MPG, breaks=seq(min(simSD_MPG), max(simSD_MPG) + 0.1, 0.1), probability=TRUE)
```



The bootstrap standard error is:

```
[1] 0.6111879
```

The mean of our boostrap distribution is

```
[1] 4.256919
```

Thus we see that our bootstrap distribution for the standard deviation is slightly skewed to the left and it has a mean of about 4.3, which is close to the standard deviation of our original sample. Additionally, our bootstrap distribution for standard deviation has a standard error of about 0.6.

e. Use the bootstrap samples in part (d) to construct a 95% boostrap t-confidence interval for the true standard deviation of the vehicle mpg. Would it be appropriate to give such a bootstrap t-confidence interval? Explain your answer.

We can run the following code in R to get a 95% confidence interval for the average standard deviation of the vehicle:

```r
lower = mean(simSD_MPG) - abs(qt(.975, length(simSD_MPG) - 1) * sd(simSD_MPG))
upper = mean(simSD_MPG) + abs(qt(.975, length(simSD_MPG) - 1) * sd(simSD_MPG))
c(lower, upper)
```

```
[1] 3.057594 5.456244
```

We can test this confidence interval by running the following code in R:

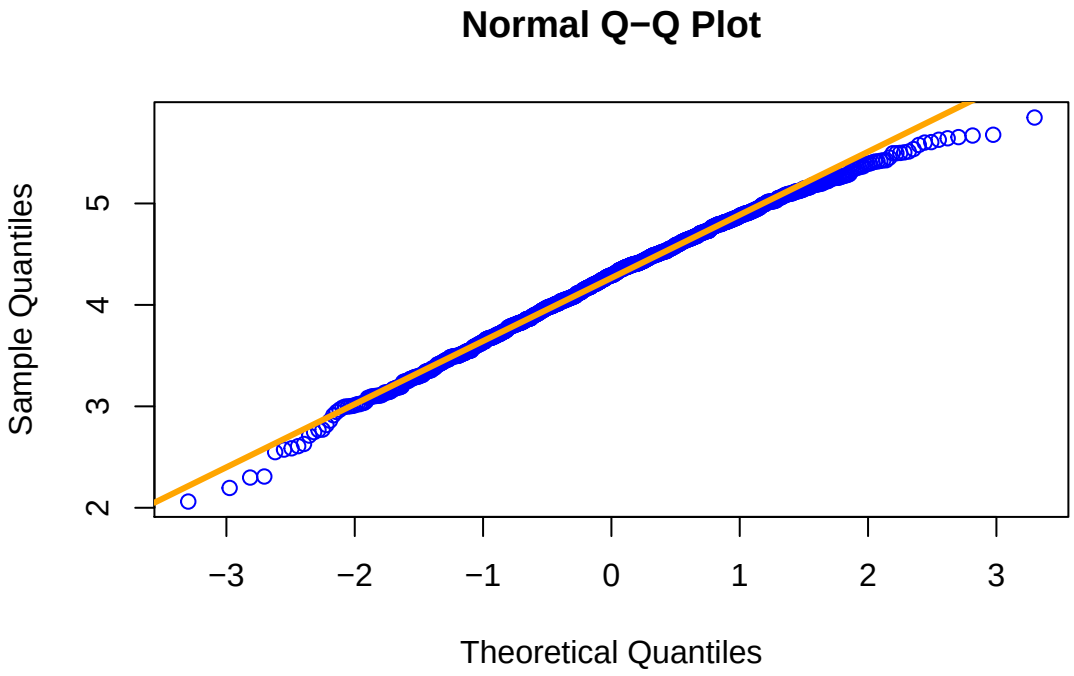
```
length( simSD_MPG[ simSD_MPG >= lower & simSD_MPG <= upper]) / length(simSD_MPG)
```

```
[1] 0.9570312
```

Thus we see that we have accurately created a 95% confidence interval for the bootstrapped standard deviation.

To check if it is appropriate to report our confidence interval, we should check the normality of our bootstrapped standard deviation distribution. We do this by looking at a normal quantile plot with R:

```
qqnorm(simSD_MPG, col="blue")
qqline(simSD_MPG, col="orange", lwd=3)
```



This looks almost normal, but the left side of the plot seems to deviate slightly from our normal quantile line. Let's run a Shapiro-Wilk test to see if our P-value is within a 5% threshold:

```
    Shapiro-Wilk normality test

data:  simSD_MPG
W = 0.99459, p-value = 0.0009837
```

It looks like our P-value is usually within our threshold (Although it seems to vary by quite a bit). This means that our plot is not normally distributed, i.e. that it would not be appropriate for us to report our 95% bootstrap t-confidence interval for the true standard deviation of the mpg.

f. Use the bootstrap samples in part (d) to construct a 95% bootstrap percentile confidence interval for the true standard deviation of the vehicle mpg.

We can use R to construct the 95% bootstrap percentile confidence interval for the true standard deviation of the mpg by running the following code:

```
c(quantile(simSD_MPG, 0.025), quantile(simSD_MPG, 0.975))
```

```
    2.5%    97.5%
3.027077 5.364453
```

Problem 2 (21 pts)

In 2002, California Polytechnic State University (San Luis Obispo) conducted a survey to collect data to investigate the question of whether back aches might be due to carrying heavy backpacks. The variables 'BackPackWeight,' 'BackProblems,' and 'Gender' are considered for this problem.

```
'data.frame':   100 obs. of  3 variables:
 $ BackpackWeight: int  9 8 10 6 8 5 8 4 5 2 ...
 $ BackProblems  : int  1 0 1 0 0 0 0 1 0 0 ...
 $ Gender        : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 2 2 1 1 1 ...
```

a. Let μ_0 be the true mean weight of the backpacks carried by students without an aching back problem, and let μ_1 be the true mean weight of backpacks carried by students with an aching back problem. Use Student's t-test and the liberal degree of freedom, at significance level $\alpha = 5\%$, to examine whether the students with an aching back problem tend to carry heavier backpacks than the students without an aching back problem. Report the standard error used in this t-test.

We first state our hypotheses:

$$H_0 : \mu_0 = \mu_1,$$
$$H_A : \mu_0 < \mu_1.$$

To run the Student's t-test to examine whether students with an aching back tend to carry heavier backpacks than students without an aching back problem we can run the following code in R:

```
# Note: we specify var.equal=TRUE to force liberal degree of freedom
t.test(BackpackWeight ~ BackProblems, alternative="less", var.equal=TRUE, paired=FALSE)
```

```
    Two Sample t-test

data:  BackpackWeight by BackProblems
t = -1.1879, df = 98, p-value = 0.1189
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.5829379
sample estimates:
mean in group 0 mean in group 1
       11.19118        12.65625
```

We see that our Student's t-test yields a P-value of about 0.24, which is much larger than our significance level $\alpha = 5\%$. Thus we do not have sufficient evidence to reject the null hypothesis that the true means μ_0 and μ_1 are equal. In other words, we do not have sufficient evidence to report a correlation between back pain and backpack weight.

b. Now use the permutation test with 1024 replications to examine whether the students with an aching back problem tend to carry heavier backpacks than the students without an aching back problem. Be sure to address each of the following questions in completing this

problem: i) Make a histogram of the permuation distribution and indicate the portion under the histogram for computing the P-value of the permuation test. ii) Report the P-value of the permuation test. iii) Draw an appropriate conclusion at significance level of 5%.

First we state our hypotheses:

$$H_0 : \mu_0 = \mu_1,$$
$$H_A : \mu_0 < \mu_1.$$

c. Use the chi-square test at significance level of 5% to examine whether 'Gender' and 'BackProblems' are independent. Compute the chi-square statistic manually to explain your answer.

Our null hypothesis is that 'Gender' and 'BackProblems' are independent attributes, and our alternative hypothesis is that they are dependent attributes.

We run the following code in R to generate a contingency table for our data:

```
# Splice the data for our contingency table
pain_m = backPack[ which(BackProblems == 1 & Gender == 'Male'), ]
pain_f = backPack[ which(BackProblems == 1 & Gender == 'Female'), ]
no_pain_m = backPack[ which(BackProblems == 0 & Gender == 'Male'), ]
no_pain_f = backPack[ which(BackProblems == 0 & Gender == 'Female'), ]

# Calculate the totals (num_pain + num_no_pain and num_m + num_f should both be 22)
num_pain = nrow(pain_m) + nrow(pain_f)
num_no_pain = nrow(no_pain_m) + nrow(no_pain_f)
num_m = nrow(backPack[ which(Gender == 'Male'), ])
num_f = nrow(backPack[ which(Gender == 'Female'), ])

obs = c(nrow(pain_m), nrow(pain_f), nrow(no_pain_m), nrow(no_pain_f)) # I'll use this later

# Create the table
makeTable = function(value, totals) {
  # Must have exactly 1 degree of freedom
  # Value is the top left value in the table
  # Totals are the sum of row elements in decending order, then the columns in ascending order.
  v2 = totals[1] - value
  v3 = totals[3] - value
  v4 = totals[2] - v3
  ans = matrix(c(value, v2, totals[1],
                 v3, v4, totals[2], totals[3],
                 totals[4], totals[1] + totals[2]),
               ncol=3)
  colnames(ans) = c("Male", "Female", "Total")
  rownames(ans) = c("Pain", "No Pain", "Total")
  ans
}

totals = c(num_pain, num_no_pain, num_m, num_f)
as.table(makeTable(nrow(pain_m), totals))
```

```
        Male Female Total
Pain       8     37    45
No Pain   24     31    55
Total     32     68   100
```

Now we run the following code in R to compute the expected frequencies:

```
total = num_m + num_f
row1 = c(num_pain * num_m / total, num_pain * num_f / total)
row2 = c(num_no_pain * num_m / total, num_no_pain * num_f / total)
exp = c(row1, row2) # I'll use this later
exp_table = matrix(c(row1, row2), ncol=2)
colnames(exp_table) = c("Male", "Female")
rownames(exp_table) = c("Pain", "No Pain")
as.table(exp_table)
```

```
        Male Female
Pain    14.4   30.6
No Pain 17.6   37.4
```

We now run the following code in R to compute our chi-square statistic:

```
chi_stat = sum(((obs - exp)^2) / exp)
chi_stat
```

```
[1] 7.605466
```

Now we use this statistic to compute our P-value:

```
1 - pchisq(chi_stat, 1)
```

```
[1] 0.005819161
```

We see that our P-value is much smaller than 5%, so we have sufficient evidence to reject our null hypothesis that 'Gender' and 'BackProblems' are independent. This means that we have shown that our data supports the alternative hypothesis that 'Gender' and 'BackProblems' are dependent attributes.

d. In part (c), if the alternative test is "female students tend to have a higher chance of having back problems than male students," what would be the P-value of the chi-square test? Explain your answer.

e. Use Fisher's exact test at significance level of 5% to examine whether 'Gender' and 'BackProblems' are independent. Report how many contingency tables that are at least as extreme as the observed table, but do not list all the tables.

f. Use the permuation test with 1024 replications at significance level of 5% to test the alternative hypothesis: "female students tend to have a higher chance of having back problems than male students."

Problem 3 (8 pts)

A chain moves on the set $S = \{1, 2, 3, 4, 5, 6\}$ according to the following rule. On each successive step, the chain moves equally likely from an integer a to those integers that are coprime to a. For example, if the chain starts at 1, it will move to each of the six states equally likely (since 1 is coprime to any integer). If the chain starts at 2, if will move to 1, 3, and 5 with an equal probability of $\frac{1}{3}$.

a. Find the one-step probability transition matrix and the stationary distribution of this chain. Express your answers as fractions.

We construct the original transition matrix:

$$\mathbf{P} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

To get the stationary distribution, we need to solve the linear system $\mathbf{vP} = \mathbf{v}$ for $\mathbf{v}$. This system would be unpleasant to solve by hand, so we use R instead:

```
# Transition matrix
P = matrix( c( 1/6, 1/6, 1/6, 1/6, 1/6, 1/6,
               1/3, 0, 1/3, 0, 1/3, 0,
               1/4, 1/4, 0, 1/4, 1/4, 0,
               1/3, 0, 1/3, 0, 1/3, 0,
               1/5, 1/5, 1/5, 1/5, 0, 1/5,
               1/2, 0, 0, 0, 1/2, 0),
            nrow=6, ncol=6, byrow=TRUE)

# Function that will compute the stationary vector of any
# row-oriented transition matrix
getEquilibrium = function(transition) {
  eigen_stuff = eigen(t(transition))
  index = match(1, round(eigen_stuff$values, digits=2))
  eigen_vec1 = eigen_stuff$vectors[ , index[1]]
  library(MASS)
  fractions(abs(eigen_vec1) / sum(abs(eigen_vec1)))
}

v = getEquilibrium(P)
v
```

```
[1] 6/23 3/23 4/23 3/23 5/23 2/23
```

We can confirm that this is a stationary distribution of our transition matrix by taking the matrix product $\mathbf{vP}$:

```
fractions( v %*% P )
```

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 6/23 3/23 4/23 3/23 5/23 2/23
```

b. If the chain initially starts at 3, use R to find the 5th-step distribution of this chain. Give an interpretation of the 5th-step distribution in context.

```
Loading required package: expm

Warning: package 'expm' was built under R version 3.4.4

Loading required package: Matrix


Attaching package: 'expm'

The following object is masked from 'package:Matrix':
```

```
expm
```

We can use R to compute the 5th step distribution as follows:

```
u = matrix(c(0, 0, 1, 0, 0, 0), nrow=1, byrow=TRUE)
c(u %*% (P %^% 5) )
```

```
[1] 0.25870177 0.13450039 0.16959877 0.13450039 0.21448881 0.08820988
```

The 5th step distribution gives us some idea as to what state we expect our system to be in after 5 steps given an intial state. This 5th distribution tells us that the two most likely states for our chain to be in after 5 steps are 1 and 5. This makes sense, since 1 is coprime to every number in our set, and 5 is coprime to every number in our set except itself. Thus our chain moves most easily to 1, followed by 5. We also notice that 2 and 4 have the same probability, which makes sense in our context because 4 is simply the square of 2, i.e. it has the same factors. Thus everything in our set that is coprime to 2 is also coprime with 4, and vise-versa. The number 6 has factors of 2 and 3, so it cannot be reached directly from 2, 3, and 4. Thus it has the lowest probability.

Problem 4 (6 pts)

Consider the following process. Bob has two coins, one of which is fair, and the other of which has heads on both sides. He gives these two coins to Alice, who chooses one of them at random (each with probability 50%). During the rest of the process, Alice uses only the coin that she chose. She now proceeds to toss the coin many times, reporting the results. We consider this process to consist solely of what she reports to us.

a. Given that she reports a head on the jth toss, what is the probability that a head is thrown on the (j + 1)th toss? [Hint: Bayes' rule]

Let 'A' denote the event of getting a head on the jth toss, and let 'B' denote the event of getting a head on the (j+1)th toss. The probability that Alice will get a head on the (j+1)th toss given that she tossed a head on the jth toss is given by:

$$Pr(B|A) = \frac{Pr(A \cap B}{Pr(A)}.$$

Since 'A' and 'B' are independent events, this becomes

$$Pr(B|A) = \frac{Pr(A)Pr(B)}{Pr(A)} = Pr(B).$$

The probability of event 'B' happening depends on the original event of Alice picking a coin. Since one coin has a 100% chance of yielding heads, the other 50%, and Alice picked between them with equal probability,

$$Pr(B) = \frac{1}{2}(1) + \frac{1}{2}\left(\frac{1}{2}\right) = 0.75.$$

b. Now assume that the process is in a state of "heads" on both the jth toss and the (j + 1)th toss. Find the probability that a head comes up on the (j + 2)th toss. [Hint: Bayes' rule]

c. Is this process a Markov chain? Explain your answer.