hw2

Joshua Burkhart

Homework 2: Probability

Part 1: Hot Hands

Getting Started

```
download.file(
  "http://www.openintro.org/stat/data/kobe.RData",
  destfile ="kobe.RData")
load("kobe.RData")
head(kobe)
```

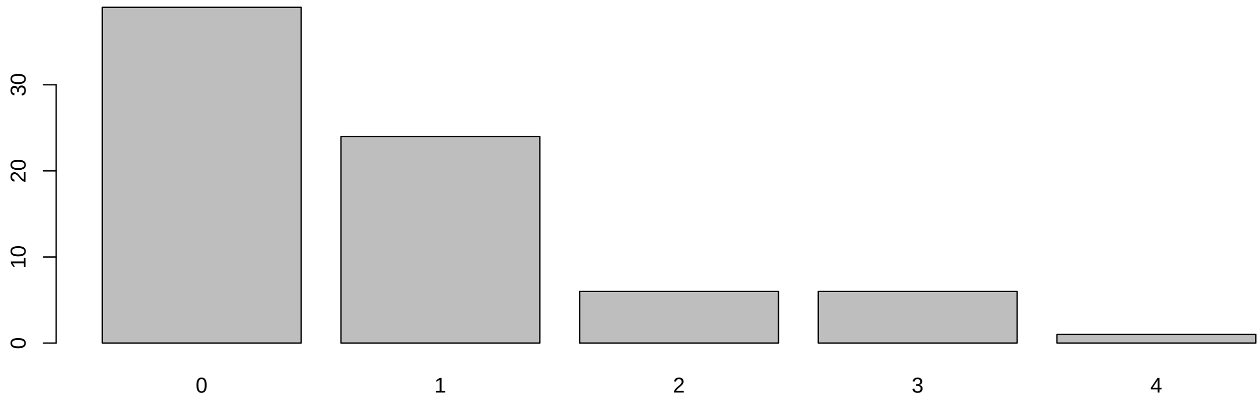
```
   vs game quarter time
1 ORL    1       1 9:47
2 ORL    1       1 9:07
3 ORL    1       1 8:11
4 ORL    1       1 7:41
5 ORL    1       1 7:03
6 ORL    1       1 6:01
                                              description basket
1                 Kobe Bryant makes 4-foot two point shot      H
2                               Kobe Bryant misses jumper      M
3                        Kobe Bryant misses 7-foot jumper      M
4 Kobe Bryant makes 16-foot jumper (Derek Fisher assists)      H
5                         Kobe Bryant makes driving layup      H
6                               Kobe Bryant misses jumper      M
```

```
kobe$basket[1:9]
```

```
[1] "H" "M" "M" "H" "H" "M" "M" "M" "M"
```

1. What does a streak length of 1 mean, i.e. how many hits and misses are in a streak of 1? What about a streak length of 0? > One hit and one miss are in a streak of one. Zero hits and one miss are in a streak of zero.

```
kobe_streak <- calc_streak(kobe$basket)
barplot(table(kobe_streak))
```



2. Describe the distribution of Kobe's streak lengths from the 2009 NBA finals. What was his typical streak length? How long was his longest streak of baskets? > Distribution skews up (positive). Most common streak length was zero. Longest streak length was 4

Simulations in R

```
outcomes <- c("heads", "tails")
sample(outcomes, size = 1, replace = TRUE)
```

```
[1] "heads"
```

```
sim_fair_coin <- sample(outcomes, size = 100, replace = TRUE)
```

```
sim_fair_coin
```

```
  [1] "tails" "heads" "heads" "heads" "heads" "tails" "tails" "heads"
  [9] "tails" "heads" "heads" "heads" "tails" "heads" "tails" "heads"
 [17] "tails" "heads" "heads" "heads" "heads" "heads" "heads" "heads"
 [25] "heads" "heads" "heads" "heads" "tails" "tails" "tails" "heads"
 [33] "tails" "heads" "tails" "heads" "heads" "heads" "heads" "heads"
 [41] "heads" "heads" "heads" "heads" "tails" "heads" "tails" "heads"
 [49] "tails" "heads" "heads" "heads" "tails" "tails" "tails" "tails"
 [57] "tails" "tails" "tails" "heads" "tails" "heads" "tails" "heads"
 [65] "heads" "tails" "tails" "tails" "tails" "heads" "tails" "tails"
 [73] "heads" "heads" "tails" "heads" "tails" "heads" "heads" "tails"
 [81] "tails" "tails" "heads" "tails" "heads" "heads" "tails" "heads"
 [89] "heads" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
 [97] "tails" "tails" "heads" "tails"
```

```
table(sim_fair_coin)
```

```
sim_fair_coin
heads tails
   53    47
```

```
sim_unfair_coin <- sample(outcomes,
                          size = 100,
                          replace = TRUE,
                          prob = c(0.2, 0.8))
```

```
sim_unfair_coin
```

```
  [1] "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
  [9] "tails" "heads" "heads" "tails" "heads" "heads" "tails" "tails"
 [17] "heads" "tails" "tails" "tails" "tails" "tails" "tails" "heads"
 [25] "tails" "tails" "tails" "tails" "heads" "tails" "tails" "tails"
 [33] "tails" "heads" "tails" "tails" "heads" "heads" "tails" "tails"
 [41] "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
 [49] "tails" "tails" "heads" "tails" "tails" "tails" "tails" "tails"
 [57] "heads" "tails" "heads" "tails" "heads" "tails" "tails" "tails"
 [65] "tails" "heads" "tails" "tails" "tails" "tails" "heads" "tails"
 [73] "tails" "tails" "tails" "heads" "tails" "heads" "tails" "tails"
 [81] "tails" "tails" "tails" "tails" "tails" "heads" "tails" "tails"
 [89] "tails" "tails" "heads" "heads" "heads" "tails" "tails" "tails"
 [97] "tails" "tails" "tails" "heads"
```

```
table(sim_unfair_coin)
```

```
sim_unfair_coin
heads tails
   23    77
```

3. In your simulation of flipping the unfair coin 100 times, how many flips came up heads? > 18

```
?sample
```

Simulating the Independent Shooter

```
outcomes <- c("H", "M")
sim_basket <- sample(outcomes, size = 1, replace = TRUE)
```

4. What change needs to be made to the sample function so that it reflects a shooting percentage of 45%? Make this adjustment, then run a simulation to sample 133 shots. Assign the output of this simulation to a new object called sim_basket. > The .2, .8 probabilities should be changed to .45, .55 and the size should be changed to 133.

```
outcomes <- c("H", "M")
sim_basket <- sample(outcomes,
                     size = 133,
                     replace = TRUE,
                     prob = c(0.45, 0.55))
```

```
kobe$basket
```

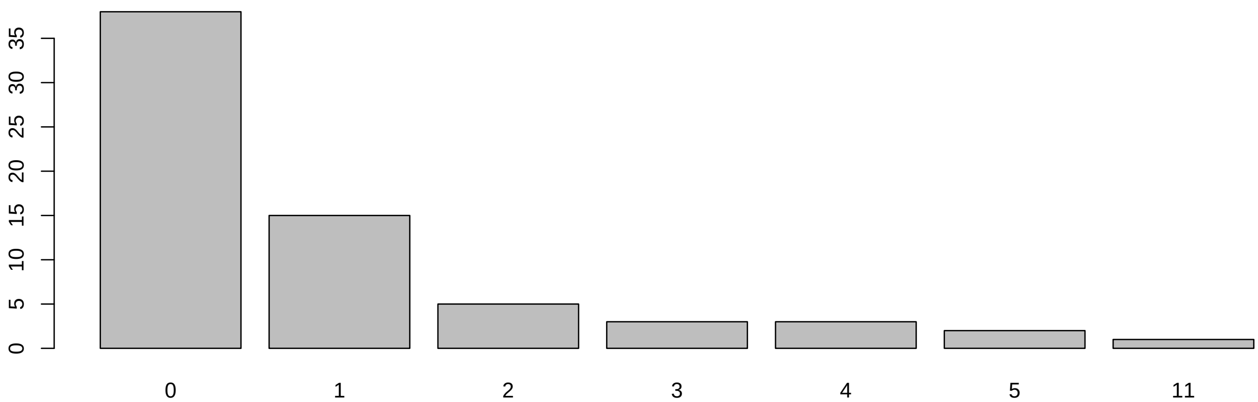
```
  [1] "H" "M" "M" "H" "H" "M" "M" "M" "M" "H" "H" "H" "M" "H" "H" "M" "M"
 [18] "H" "H" "H" "M" "M" "H" "M" "H" "H" "H" "M" "M" "M" "M" "M" "M" "H"
 [35] "M" "H" "M" "M" "H" "H" "H" "H" "M" "H" "M" "M" "H" "M" "M" "H" "M"
 [52] "M" "H" "M" "H" "H" "M" "M" "H" "M" "H" "H" "M" "H" "M" "M" "M" "H"
 [69] "M" "M" "M" "M" "H" "M" "H" "M" "M" "H" "M" "M" "H" "H" "M" "M" "M"
 [86] "M" "H" "H" "H" "M" "M" "H" "M" "M" "H" "M" "H" "H" "M" "H" "M" "M"
[103] "H" "M" "M" "M" "H" "M" "H" "H" "H" "M" "H" "H" "H" "M" "H" "M" "H"
[120] "M" "M" "M" "M" "M" "M" "H" "M" "H" "M" "M" "M" "M" "H"
```

```r
sim_basket
```

```
  [1] "M" "H" "H" "H" "H" "H" "M" "M" "M" "H" "H" "H" "M" "M" "H" "M" "M"
 [18] "M" "M" "M" "M" "M" "H" "M" "H" "H" "M" "M" "H" "M" "H" "H" "M" "H"
 [35] "M" "H" "M" "H" "M" "H" "M" "M" "H" "H" "M" "M" "M" "H" "H" "H" "H"
 [52] "H" "H" "H" "H" "H" "H" "H" "M" "M" "H" "M" "H" "H" "H" "M" "H" "M"
 [69] "M" "M" "H" "H" "H" "M" "M" "M" "M" "M" "H" "H" "H" "H" "H" "M" "M"
 [86] "M" "H" "M" "M" "H" "H" "H" "H" "M" "H" "H" "M" "M" "H" "M" "M" "H"
[103] "M" "M" "M" "M" "H" "H" "H" "H" "M" "H" "H" "H" "H" "M" "H" "H" "M"
[120] "H" "M" "M" "M" "M" "H" "M" "M" "H" "M" "M" "M" "M" "M"
```

On your Own Comparing Kobe Bryant to the Independent Shooter

```r
kobe_streak <- calc_streak(sim_basket)
barplot(table(kobe_streak))
```




```r
table(kobe_streak)
```

```
kobe_streak
 0  1  2  3  4  5 11
38 15  5  3  3  2  1
```

Describe the distribution of streak lengths. What is the typical streak length for this simulated independent shooter with a 45% shooting percentage? How long is the player's longest streak of baskets in 133 shots? > Distribution skews up (positive) as it did with actual Kobe data. Most common streak length was zero. Longest streak length was 5.

If you were to run the simulation of the independent shooter a second time, how would you expect its streak distribution to compare to the distribution from the question above? Exactly the same? Somewhat similar? Totally different? Explain your reasoning. > I would expect it to be similar to the above distribution as it was generated using the same paremeters, i.e each basket was simulated independently using a .45, .55 probability. I would find it odd if the distribution were exactly the same and would be very surprised if it were totally different.

How does Kobe Bryant's distribution of streak lengths compare to the distribution of streak lengths for the simulated shooter? Using this comparison, do you have evidence that the hot hand model fits Kobe's shooting patterns? Explain. > The simulated distribution is similar to Kobe's. It skews the same direction and has a similar number of streaks of each length. The hot hand model would predict Kobe's distribution would appear different from the independent simulation. As they are similar, this would not qualify as evidence that Kobe follows such a model.

To turn in

Part 2: Probability Exercises

To turn in You can turn in your answers to Part 2 with your answers to Part 1 if you choose to use Latex within your R markdown file.

1. A coin is tossed three times and the sequence of heads and tails is recorded.

a. list the sample space.

$$\{H,H,H\}$$
$$\{H,H,T\}$$
$$\{H,T,H\}$$
$$\{H,T,T\}$$
$$\{T,H,H\}$$
$$\{T,H,T\}$$
$$\{T,T,H\}$$
$$\{T,T,T\}$$

b. list the elements that make up the following events:

(1) A = at least two heads;

$$\{H,H,H\}$$
$$\{H,H,T\}$$
$$\{H,T,H\}$$
$$\{T,H,H\}$$

```
(2) B = the first two tosses are heads;
```

$$\{H,H,H\}$$
$$\{H,H,T\}$$

```
(3) C = the last toss is a tail.
```

$$\{H,H,T\}$$
$$\{H,T,T\}$$
$$\{T,H,T\}$$
$$\{T,T,T\}$$

c. List the elements of the following events:

   (1) complement of A,

$$\{T, T, T\}$$
$$\{T, T, H\}$$
$$\{T, H, T\}$$
$$\{H, T, T\}$$

`(2) A   B,`

$$\{H, H, H\}$$
$$\{H, H, T\}$$

`(3) A   C.`

$$\{H, H, H\}$$
$$\{H, H, T\}$$
$$\{H, T, H\}$$
$$\{T, H, H\}$$
$$\{H, T, T\}$$
$$\{T, H, T\}$$
$$\{T, T, T\}$$

2. In a city, 65% of people drink coffee, 50% drink tea, and 25% drink both. What is the probability that a person chosen at random will drink at least one of coffee or tea? Will drink neither?

   25% drink both coffee and tea
   25% must drink coffee only (50 - 25)
   40% must drink tea only (65 - 25)
   10% must not drink coffee or tea (100 - (25 + 25 + 40))

3. A game in a state lottery selects four numbers from the following set: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, with no number being repeated. How many possible groups of four numbers are there?

   Assuming order doesn't matter, as in Powerball
   (http://www.powerball.com/powerball/pb_howtoplay.asp)

$$\frac{n!}{m!(n-m)!} \text{ from Theorem 2.3.3 (p. 65)}$$

$$\begin{aligned} &\frac{10!}{4!(10-4)!} \\ &= \frac{10!}{4!6!} \\ &= \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} \\ &= 210 \end{aligned}$$

4. A bag of 15 apples contains 2 rotten apples. Four apples are selected at random.

a. what is the probability that none of the selected apples is rotten?

$$P(A^c) = 1 - P(A) \text{ from Property 1 (p. 58)}$$

$$\begin{aligned} 1 - (P(a1 = rotten)+ \\ P(a2 = rotten)+ \\ P(a3 = rotten)+ \\ P(a4 = rotten)) \\ = 1 - (\frac{2}{15} + \frac{2}{14} + \frac{2}{13} + \frac{2}{12}) \\ = 1 - 0.5967033 \\ = 0.4032967 \end{aligned}$$

b. what is the probability that at least one of the selected apples is rotten?

$$P(A^c) = 1 - P(A) \text{ from Property 1 (p. 58)}$$

$$\begin{aligned} 1 - 0.4032967 \text{ (result from part a., above)} \\ = 0.5967033 \end{aligned}$$

5. Give an example of three events A, B, C which are not independent, yet satisfy: P(A B C) = P(A)P(B)P(C) (Hint: consider simple and extreme cases.)

   Imagine the case where:
   A = I am my iPhone
   B = I am designed by Apple in California
   C = I am assembled in China

$$\begin{aligned} P(ABC) &= P(A)P(B)P(C) \\ &= 0 \times\ .99 \times\ .99 \\ &= 0 \end{aligned}$$

6. A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase "free money" is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention "free money". What is the probability that it is spam?

   I = email is spam
   N = email not spam F = email contains 'free money'

$$\begin{aligned} P(I) = .8 \\ P(N) = 1 - P(I) = .2 \\ P(F|I) = .1 \\ P(F|N) = .01 \\ P(F) = P(N)P(F|N) + P(I)P(F|I) \text{ from Law of Total Prob.} \\ = .2 \times .01 + .8 \times .1 \\ = .082 \end{aligned}$$

$$
\begin{aligned}
P(I|F) &= \frac{P(F|I)P(I)}{P(F)} \text{ from Bayes Rule} \\
&= \frac{.1 \times .8}{.082} \\
&= 0.9756098
\end{aligned}
$$

7. Propose a probabilty question of your own! The requirements for a good problem are: A clearly stated question about random process.

   The general risk of heart attack during one's lifetime is .2. Genotype 'AG' (mutant) indicates one will have a heart attack sometime in one's life with a probability of .56 and that one cannot be revived. The 'AG' mutant appears in 7% of the population. You have not been genotyped and are experiencing a heart attack. What is the probability you have a normal, i.e. non-mutant 'AG' genotype and, therefore, have a chance at surviving?

A clearly defined sample space.

$$
\begin{aligned}
&\{\text{mutant genotype, have a heart attack in lifetime}\} \\
&\{\text{mutant genotype, don't have a heart attack in lifetime}\} \\
&\{\text{normal genotype, have a heart attack in lifetime}\} \\
&\{\text{normal genotype, don't have a heart attack in lifetime}\}
\end{aligned}
$$

A probability function that satisfies the "Axiomatic Definiton of Probability" (p.57 of the textbook) Enough information provided to actually answer the question. A worked out solution!

H = heart attack
N = normal genotype
M = mutant genotype 'AG'

$$
\begin{aligned}
P(M) &= .07 \\
P(N) &= .93 \\
P(H) &= .2 \\
P(H|M) &= .56
\end{aligned}
$$

we'd like to find P(N | H)

$$
P(N|H) = \frac{P(H|N)P(N)}{P(H)}
$$

but we don't know P(H | N) so we use

$$
P(A^c) = 1 - P(A) \text{ from Property 1 (p. 58)}
$$

and find the probability we are a mutant

$$
\begin{aligned}
P(N|H) &= 1 - P(M|H) \\
1 - P(M|H) &= \frac{P(H|M)P(M)}{P(H)} \\
&= 1 - \frac{.56 \times .07}{.2} \\
&= 1 - 0.196 \\
&= 0.804
\end{aligned}
$$

If we really like the question then we'll put it on the midterm! We really like interesting math and amusing narratives!