Stat 201: HW5

Due: 11:59pm on Tuesday, April 7

Exercise 1: Textbook practice on observational studies and randomized experiments

1a. Exercise 11.22

a. Observational study
b. Prospective
c. Disabled women over the age of 65
d. The difference in depression rates between those with and without B12 deficiency
e. Because it is observational, we cannot make a causal conclusion

1b. Exercise 11.26

a. Observational study
b. Retrospective; they investigated the medical histories of the people who lived nearby
c. Women who lived near the accident and were under 40 at that time
d. The risk of breast cancer
e. We cannot make a causal conclusion because there was no random assignment

1c. Exercise 11.33

a. Experiment
b. Athletes who injured their hamstrings
c. The type of recovery program
d. Two different treatments (recovery programs)
e. The average time to recover/resume sports
f. Randomized
g. Not blind nor double blind
h. We can conclude causally that athletes recover faster with static stretching

1d. Exercise 11.43

a. Ensure that the runners don't know which shoes they are wearing

b. The results apply to olympic runners, not all runners

Exercise 2: Textbook practice on probability

2a. Exercise 12.10

a. $\Omega = \{2, 3, 4, 5, 6, \cdots, 12\}$

Not all simple events are equally likely.

b. $\Omega = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$

All simple events are equally likely.

c. $\Omega = \{0, 1, 2, 3, 4\}$

Not all simple events are equally likely.

d. $\Omega = \{0, 1, \cdots 10\}$

Not all simple events are equally likely.

2b. Exercise 12.22

It fails to remove the cases where a home has both a home and a pool; it misapplies the Princple of Inclusion/Exclusion.

2c. Exercise 12.48

a. 0.147
b. 0.599
c. 0.343

2d. Textbook problem 13.36

a. $P(Can \mid Mex) = \dfrac{P(Can \cap Mex)}{P(Mex)} = 0.444$

b. No because $P(Can \cap Mex) \neq 0$

c. No because $P(Can \mid Mex) \neq P(Can)$

2e. Exercise 13.58

Let A, B, C be the events corresponding to the connector coming from factories A, B, C respectively. These events partition our sample space; that is, $P(A) + P(B) + P(C) = 1$.

Let E be the event that the component is defective.

We are interested in $P(A|E)$.

Our priors: We have $P(A) = 0.7, P(B) = 0.2, P(C) = 0.1$ and $P(E \mid A) = 0.01, P(E \mid B) = 0.02, P(E \mid C) = 0.04$

We have $P(E) = P(E \mid A)P(A) + P(E \mid B)P(B) + P(E \mid C)P(C)$

$P(E) = (0.01)(0.7) + (0.02)(0.2) + (0.04)(0.1) = 0.015$

Now, we have $P(E)$ and $P(A)$, so we apply Bayes's rule:

$P(A \mid E) = P(E \mid A)\dfrac{P(A)}{P(E)} = 0.01\dfrac{0.7}{0.015} = \boxed{0.46666667}$

Exercise 3. Advanced probability tree practice

Imagine that you have a jar filled with 1,000 coins. 999 of these coins are fair coins (has "heads" on one side and "tails" on the other). The remaining coin has "heads" on both sides. You draw one coin from the jar at random. You proceed to flip the coin 10 times, and you see 10 heads in a row! You get ready to flip this selected coin for an 11th time: what is the probability that you get "heads"?

A = the next flip is heads
B = the coin is fair
C = the coin generates ten heads in a row

$$
\begin{aligned}
P(C \mid B) &= (0.5)^{10} \\
P(C \mid \neg B) &= 1 \\
P(B) &= 0.999 \\
P(\neg B) &= 0.001 \\
P(C) &= (0.5)^{10} \cdot 0.999 + 1 \cdot 0.001 = 0.00197559 \\
P(B \mid C) &= \frac{(0.5)^{10} \cdot 0.999}{0.00197559} \\
P(\neg B \mid C) &= \frac{0.001}{0.00197559} \\
P(A \mid C) &= 0.5 \cdot 0.49382004 + 1 \cdot 0.5061779 \\
&\approx \boxed{0.753}
\end{aligned}
$$

Exercise 4. Rules of random variables

Suppose that you have a "spinner" that lands on the color blue with probability 0.3, lands on the color red with probability 0.25, and lands on the color white otherwise.

There are two games that you are going to play using this spinner.

In Game A, you pay $10 to play the game. You then spin the spinner 7 times in a row. You win $1,000 if the spinner lands on blue all 7 times. Your random variable X is your profit from this game.

In Game B, you win $5 just for signing up to play. However, if you spin the spinner 3 times in a row and they all land on red, then you need to pay $500 (and you don't get your original $5). Your random variable Y is your profit from this game.

4a. What is the distribution, expected value, and standard deviation of X?

The possible values of X are 990 and 10. $P(X = 990) = 0.0002187$ and $P(X = -10) = 0.9997813$

$$E[X] = 0.9997813 \times (-10) + 0.0002187 \times (990) = \boxed{-9.7813}$$

$$StDev(X) = \sqrt{(-10 + 7.813)^2(1 - 0.3^7) + 0.3^7(990 + 7.813)^2} = \boxed{14.9173}$$

4b. What is the distribution, expected value, and standard deviation of Y?

The possible values of Y are 5 and -500. $P(Y = 5) = 0.984375$ and $P(Y = -500) = 0.015625$

$$E[Y] = 0.015625 \times (-500) + 0.984375 \times (5) = \boxed{-2.890625}$$

$$StDev(Y) = \sqrt{(5 + 2.890625)^2(1 - 0.25^3) + 0.25^3(-500 - 2.890625)^2} = \boxed{62.6299}$$

4c. You play Game A and Game B. Let your total profit be denoted with Z. What is the distribution, expected value, and standard deviation of Z?

$$Z \in \{-5, -510, 995, 490\}$$
$$P(X = 990) = 0.0002187, \quad P(X = -10) = 0.9997813$$
$$P(Y = -500) = 0.015625, \quad P(Y = 5) = 0.984375$$

$$P(Z = -5) = P(X = -10, Y = 5) = (0.9997813)(0.984375) \approx 0.98416$$

$$P(Z = -510) = P(X = -10, Y = -500) = (0.9997813)(0.015625) \approx 0.0156$$

$$P(Z = 995) = P(X = 990, Y = 5) = (0.0002187)(0.984375) \approx 0.00022$$

$$P(Z = 490) = P(X = 990, Y = -500) = (0.0002187)(0.015625) \approx 0.000003$$

$$E[Z] = (-510)(0.01562) + (-5)(0.98416) + (490)(0.00000342) + (995)(0.0002153)$$
$$E[Z] \approx -7.968 - 4.9208 + 0.001675 + 0.2139 \approx \boxed{-12.67}$$

$$E[Z^2] = (-510)^2(0.01562) + (-5)^2(0.98416) + (490)^2(0.00000342) + (995)^2(0.0002153)$$
$$E[Z^2] \approx 4056.21 + 24.60 + 0.819 + 212.34 \approx 4293.97$$

$$Var(Z) = E[Z^2] - (E[Z])^2 \approx 4293.97 - (-12.67)^2 \approx 4293.97 - 160.55 \approx 4133.42$$

$$StDev(Z) \approx \sqrt{4133.42} \approx \boxed{64.3}$$

]

4d. You have the option of playing Game A once, or playing it 10 times in a row. Which option would you pick, and why?

Once, because it is a negative EV (expected value) decision. I would prefer to play zero times.

4e. You have the option of playing Game B once, or playing it 10 times in a row. Which option would you pick, and why?

Once, because it is a negative EV decision. I would prefer to play zero times.

Exercise 5: The Hot Hand in Baksetball

This problem is modified from a lab from the OpenIntro textbook.

Basketball players who make several baskets in succession are described as having a *hot hand*. Fans and players have long believed in the hot hand phenomenon; when a player makes a few baskets in a row, they become more likely to make the next basket. However, a 1985 paper by Gilovich, Vallone, and Tversky collected evidence that contradicted this belief and showed that successive shots are independent events. This paper started a great controversy that continues to this day, as you can see by Googling *hot hand basketball*.

We have a data set on Kobe Bryant's shots during the 2009 NBA Finals, where he was named MVP and where many spectators commented on how he appeared to show a hot hand. Run the following chunk to download the data:

```
kobe_basket <- read.delim("http://anna-neufeld.github.io/Stat311/oiLabs/Week4/kobe_basket.csv", head
```

The column `shot` stores all of Kobe's shots from these playoffs in order: an `H` means that he made the shot (hit) and an `M` means that he missed the shot.

We will compare Kobe Bryant's real data to simulated data from an *independent shooter* (i.e. a shooter that definitely does not have a hot hand). We will see if Kobe's data seems extreme or unusual relative to the independent shooter: if so, this would provide evidence that Kobe really did have a "hot hand".

5a. Overall, what proportion of Kobe's baskets in these playoffs were "hits"?

$$\frac{58}{133} = \boxed{43.6\%}$$

5b. Simulating an independent shooter.

The following code simulates a 10 shots from an independent random shooter whose overall "hit" probability is 0.2.

```
sample(c("H", "M"), size=10, replace=T, prob=c(0.2,0.8))
```

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

Run the code a few times to see if you can figure out what it is doing. Then, write modified code that simulates 133 shots from an independent shooter who has the same overall "hit" probability as Kobe. Why did I choose 133? If you would like to get the same answer every time you run your chunk, you can set a random seed.

There were 58 hits out of 133 shots, so the overall probability of a hit is 0.43609023, so we have

```
set.seed(120)
sample(c("H", "M"), size=133, replace=T, prob=c(0.43609023,0.56390977))
```

```
##   [1] "M" "M" "M" "H" "M" "M" "H" "M" "H" "H" "H" "H" "M" "M" "H" "M" "M" "M"
##  [19] "H" "M" "H" "M" "H" "H" "H" "M" "M" "M" "M" "M" "M" "M" "M" "H" "H" "M" "M"
##  [37] "H" "M" "M" "H" "H" "M" "M" "H" "M" "H" "M" "H" "H" "M" "M" "H" "H" "H"
##  [55] "H" "H" "M" "M" "M" "M" "H" "M" "H" "H" "H" "M" "H" "H" "H" "H" "M" "M"
##  [73] "M" "H" "H" "H" "M" "M" "M" "M" "H" "M" "H" "M" "M" "H" "M" "H" "H" "M"
```

```
##  [91] "M" "M" "H" "M" "M" "H" "M" "M" "H" "H" "M" "M" "M" "H" "H" "H" "H" "M"
## [109] "H" "M" "M" "M" "H" "H" "H" "H" "H" "M" "M" "H" "M" "M" "H" "M" "M" "H"
## [127] "H" "M" "M" "H" "H" "M" "M"
```

You chose 133 because there are 133 shots in the dataset.

5c. Counting streaks

We now have shot data from Kobe, who may or may not have had a hot hand. We also have shot data from a random, independent shooter. This random shooter is just as good at basketball as Kobe (why?), but definitely does *not* have a hot hand (why?).

The random shooter is "as good at basketball" because it has the same overall hit probability as Kobe. It does not have a hot hand because we generated all of the values and ensured that they are independent.

We want to compare Kobe's data to data from the random shooter in terms of how long their "streaks" of hits tended to be. To see what I mean by streak, let's look at Kobe's sequence of hits and misses from the first quarter of game 1.

```
kobe_basket %>% filter(game==1, quarter==1) %>% pull(shot)
```

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

A streak is defined as the number of hits before the next miss. For our data:

H M | M | H H M | M | M | M

the streak lengths are 1,0,2,0,0,0.

Counting streak lengths manually for all 133 shots would get tedious. The code below defines a custom function that you can use to count streaks. You are not required to understand the code here, but if you are curious please ask. For your purposes: once you run the chunk below, you will be able to apply the function `calc_streak()` to any sequence of "H"s and "M"s.

```
calc_streak <- function(shots) {
  hits <- c(0, shots=="H", 0)
  misses <- which(hits==0)
  streaklengths <- diff(misses)-1
  return(data.frame(length= streaklengths))
}
```

Apply the function `calc_streak` to `kobe_basket$shot` and interpret the result.

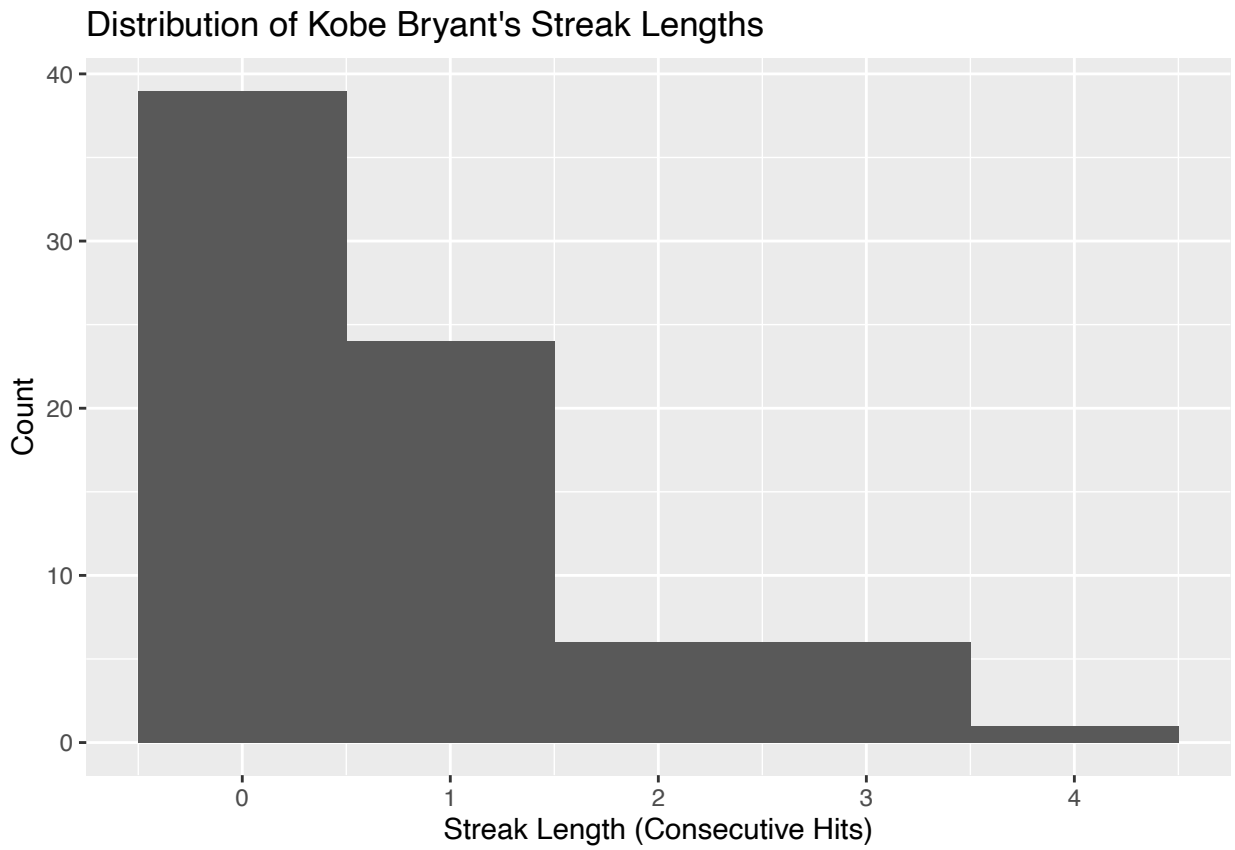
```
kobe_streaks <- calc_streak(kobe_basket$shot)
```

The result is a vector containing the length of each of Kobe's streaks, in order. Kobe had 76 streaks, the longest of which was 4 "hits" long.

5d. Describe the distribution of Kobe's streak lengths. A histogram is probably a good idea.

```r
library(tidyverse)

kobe_streaks %>%
  ggplot(aes(x = length)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = 0:max(kobe_streaks$length)) +
  labs(
    title = "Distribution of Kobe Bryant's Streak Lengths",
    x = "Streak Length (Consecutive Hits)",
    y = "Count"
  )
```



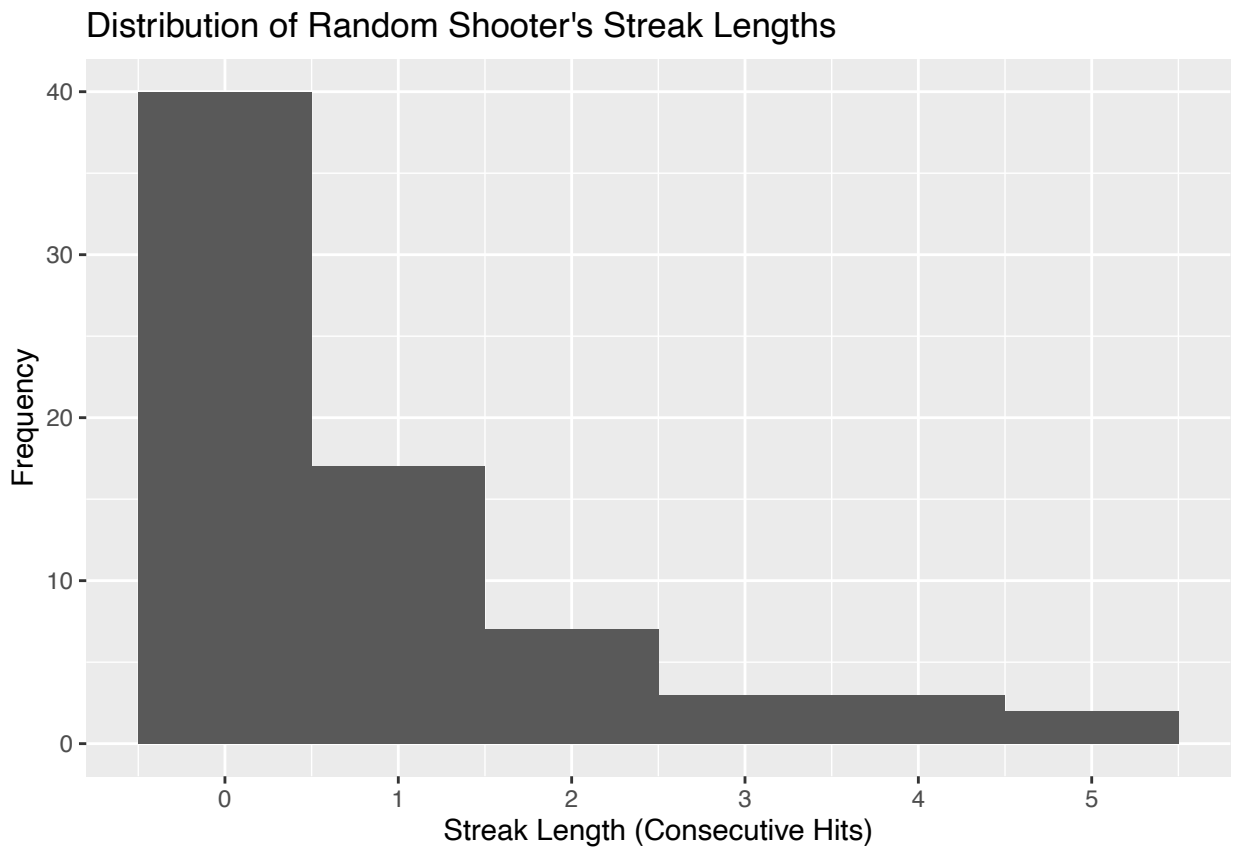
5e. Now, calculate the streak lengths for your random, independent shooter from part b. Describe the distribution of their streak lengths. A histogram is probably a good idea.

Setting a seed in 5b so that you get the same answer every time is also probably a good idea. It is also fine to re-generate a new random shooter here, with a new seed.

```r
set.seed(120)
random_shooter = sample(c("H", "M"), size=133, replace=T, prob=c(0.43609023,0.56390977))
random_streaks = calc_streak(random_shooter)

random_streaks %>%
  ggplot(aes(x = length)) +
```

```
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = 0:max(random_streaks$length)) +
  labs(
    title = "Distribution of Random Shooter's Streak Lengths",
    x = "Streak Length (Consecutive Hits)",
    y = "Frequency"
  )
```



5f. Based on the comparison between 4d and 4e, do you think that you have strong evidence that Kobe had a "hot hand" in these playoffs? Why or why not?

If Kobe had a hot hand, we should expect more long streaks than the random shooter. However, this is not the case. Kobe does not have more long streaks than the random shooter, suggesting that Kobe's streaks are consistent with chance.

```
max(kobe_streaks$length)
```

```
## [1] 4
```

```
max(random_streaks$length)
```

```
## [1] 5
```