```
#### Question 1 ####
# You roll five fair, six-sided dice. What is the probability that the sum of the five dice is 20?
# Please round your answer to four decimal places.

6^5 # 7776 total combinations
library(dice) # let's use the dice package to determine probability
diceProbs <- getSumProbs(ndicePerRoll = 5,
                         nsidesPerDie = 6,
                         nkept = 5)
diceProbs$probabilities[16,] # sum of twenty row
# Sum  Probability Ways to Roll
# 20    0.08371914     651
diceProbs$probabilities[16,2] # probability column for sum twenty

Q1 <- round(diceProbs$probabilities[16,2], 4)
Q1
# Probability of rolling a sum of 20 is 0.0837, or 8.37%!


#### Question 2 ####
# A company has invented a new device that turns red when exposed to an injured pitcher. The
# company claims that the device predicts pitcher injuries with 90% accuracy. That is, given that a
# pitcher is actually injured, the device will turn red 90% of the time. Given that the pitcher is not
# injured, the device will turn red 20% of the time. Assume that the population-level injury risk for
# pitchers is 5%.

# A randomly selected player takes the test and the device turns red. What is the probability that
# he is injured?

#                        P(+|injured)P(injured)
#  P(injured|+) = ----------------------------
#                                 P(+)
#
# P(injured) is prior probability of being injured (5% or 0.05)
# P(injured|+) is the probability of having an injury given the device turns red
# P(+|injured) is the probability of the device turning red when the pitcher is injured (90% or 0.90)
# P(+) is ? need to partition based on injury status
   # Pitcher is either injured or healthy so partition is easier to write
   # P(+) = P(+|injured)*P(injured) + P(+|healthy)*P(healthy)
     # P(+|injured) = 0.90
     # P(injured) = 0.05
     # P(+|healthy) = 0.20
     # P(healthy) = 0.95 (complement rule, since 5% chance of pitcher being injured)
partition_prob <- (0.90)*(0.05) + (0.20)*(0.95) # calculating P(+)

device_success <- ((0.90)*(0.05))/(partition_prob) # calculate device's success
Q2 <- round(device_success,3)
Q2
```

```
# Probability of the pitcher/player being injured is 0.191, or 19.1%!

#### Question 3 ####

# You are given a data set containing all 2019 batted balls with columns for Exit Velocity (in mph) and
Vertical Launch Angle (in degrees), as well as whether or not the ball was a home run. Your goal is to
predict the probability of a home run given these two variables. You start by fitting a logistic regression
model for the probability of a batted ball being a home run using only balls hit 90 mph or harder. The
following coefficients are estimated:

# Intercept: -41.2

# Exit Velocity: 0.363

# Vertical Launch Angle: 0.096


# Observed Model: HomeRun = -41.2 + 0.363(ExitVelo) + 0.096(LaunchAngle)

# What is the predicted probability of a home run, under this model, for a batted ball hit 100 mph

# at 30 degrees?

HRProp <- -41.2 + (0.363*100) + (0.096*30)

Q3 <- HRProp

Q3 # -2.02

# The probability of a home run, using this model, for this batted ball instance is 2.02% lower than an
average home run batted ball in this dataset.

# Is the marginal effect of an additional mph of exit velocity on the probability of a home run higher for
a ball hit at 105 mph and 30 degrees, or 110 mph and 25 degrees, under this model?

# test 105 and 30

-41.2 + (0.363*105) + (0.096*30) # -0.205

-41.2 + (0.363*106) + (0.096*30) # 0.158

# test 120 and 25

-41.2 + (0.363*110) + (0.096*25) # 1.13

-41.2 + (0.363*111) + (0.096*25) # 1.493

# marginal effect

# instance 1

  (-0.205-0.158)/(-0.205) # 1.77 or 177%
```

# instance 2

```
  (1.493-1.13)/(1.493) # 0.24 or 24%
```

# The marginal change of an additional mph of exit velocity on the probability of a home run is higher for a ball initially hit at 105 mph and 30 degrees than one hit at 110 mph and 25 degrees under this model.

# How do you know?

  # I know this because initially 105 and 30 (instance 1) has a probability of being a home run lower than average but then jumps above average once we increase the mph by one whereas 110 and 25 (instance 2) already has a probability of being a home run higher than average and adding the one mph does not increase its home run probability significantly. Also, this is shown in the percentage improvement in odds of being a home run where instance one's odds improve by 177% and instance 2's odds improve by 24%.

# What is one change to your modeling procedure that you think would improve predictive power over this model without collecting more data?

  # I believe including all batted ball instances will improve the model's predictive power. When the data is subsetted to only include batted balls with exit velocities 90 mph or higher we lose some home run instances. I would also explore using transformations of the Exit Velocity(EV) and Launch Angle(LA) variables. By checking different transformations such as squaring Launch Angle, we are able to see that increased LA only helps home run probability up to a point where it would then become detrimental, for example line drives/fly balls become pop-ups. The transformations help to see how EV and LA may only improve probability of a home run up to a point. We could also look at other model development such as partition or boosted tree models but developing simple logistic regression further to multiple logistic regression, with transformations, should provide us the best insight!

#### Question 4 ####

# Define a "weighted coin flip pitcher" as a pitcher who strikes out batters with rate p and walk batters with rate 1-p. This pitcher allows no batted balls, does not hit batters, does not throw wild pitches, does not pick to bases, and baserunners do not attempt steals. Assume that once the pitcher starts an inning, he will complete it.

For which value of p is the pitcher expected to throw 80% scoreless innings? Please round your answer to two decimal places. Exact and approximate answers are both acceptable.

# In order for an inning to be scoreless no more than three walks can occur in the inning (so, no more than 6 batters in an inning)

# Let's use a Monte Carlo simulation and test different probabilities

```
p <- 0.2 #example probability
```

```
set.seed(9523); sample(c("S","W"), size = 6, replace = TRUE, prob = c(p,1-p))

  # example of sample to run through in Monte Carlo simulation "S" is Strikeout & "W" is walk and p is the probability

  # result: "W" "W" "S" "W" "W" "S"

  # this inning is not scoreless

ntrials <- 1000000 # number of tiems to run through Monte Carlo

counter <- 0 # number of scoreless innings

p <- 0.59 # test/adjust probability

set.seed(9523); for (i in 1:ntrials) {

          inning <- sample(c("S","W"), size = 6, replace = TRUE, prob = c(p,1-p))

          if( sum(inning == "S") >= 3 ) { counter <- counter + 1 }

        }

counter/ntrials # 0.80672

scoreless_inning_prob <- counter/ntrials

scoreless_inning_prob

Q4 <- p

Q4 # 0.59

# The value of p for which the pitcher is expected to throw 80% scoreless innings is approximately 0.59!


#### Question 5 ####

# The Cardinals lead the division by three games over the Brewers and by four games over the Cubs. The Cubs and Cardinals play each other six more times and each play other teams three times, while the Brewers play nine remaining games against other teams.

Assuming each game's result is a 50/50 coin flip, what is the probability of a three-way tie?

Please round your answer to the nearest tenth of a percent.


# all Brewers' games are independent events

# 3 Cards and Cubs games are independent; the other 6, matchups bewtween each other, are dependent events

# Cubs vs. Cardinals   6 times
# Cubs vs. randos   3 times
# Cardinals vs. randos   3 times
# Brewers vs. randos   9 times
trials <- 1000000
ties <- 0
# let's assume win totals going into the home stretch is cubs @ 0, brewers @ 1, and cards @ 4 and use another Monte Carlo simulation
set.seed(9523); for (i in 1:trials) {
            # reset wins for each trial
            cubs <- 0
            brewers <- 1
            cardinals <- 4
            # run through final 9 game home stretch
            brew_wins <- sample(c(0,1), size = 9, replace = TRUE)
            brewers <- brewers + sum(brew_wins)
            cubs_wins <- sample(c(0,1), size = 3, replace = TRUE)
            cards_wins <- sample(c(0,1), size = 3, replace = TRUE)
            cubs_cards <- sample(c(0,1), size = 6, replace = TRUE) # 1s represent Cubs wins and 0s represent Cardinals wins
            cubs <- cubs + sum(cubs_wins) + sum(cubs_cards)
            cardinals <- cardinals + sum(cards_wins) + (6 - sum(cubs_cards))
            if ( brewers == cubs & cubs == cardinals ) { ties <- ties + 1 }
           }
ties/trials # 0.01056
Q5 <- round((ties/trials)*100, 1)
Q5
# The probability of a three-way tie when the probability is rounded to the nearest tenth of a percent is 1.1%.


#### Question 6 ####
```

# Given the following information, construct a basic model that suggests the likelihood of a successful stolen base given that the runner takes off for an empty second base:

Identities of the runner, batter, pitcher, catcher and all other fielders, game state (date, score, inning, outs, count), temperature, crowd size, runner average sprint speed, pitcher release time, catcher pop time.

Note: there is no need to find data and generate an actual model. Please just describe how you would proceed assuming you had the data listed.

# Process:

- I would begin by viewing a summary of the all of the variables listed and propose and combine any rare levels in the data set.
- I would then look at any variables and see if there are zero variance or near zero variance variables we can remove.
- Replace NA values, if there are any, using median imputation.
- Once I have added and adjusted anything needed for the variables to be more accurate in their predictions, I would start model development.
- Model Development:
    - First split the data frame into Training (70% of the total data(nrow)) and Holdout (the remaining 30%) samples
    - Start by looking at some basic logistic models and developing multiple logistic to see if there are improvements
    - I would then continue to run and test models using the caret package in R
    - I would usually run through the following models and evaluate each based on the Accuracy of the data on the Holdout
    - Models:
        - Vanilla Logistic Regression
        - Regularized Logistic Regression
        - Vanilla Partition
        - Random Forest
        - Boosted Tree
        - I will also explore Neural Networks and K Nearest Neighbors models if the previous ones are not preforming as we expect
    - Once the best model is identified, I will then refine the expand grid parameters for each model to locate the best version
    - After the final model parameters are identified, I would run the last model on the test data!
    - Note: If the model is being used to be run on a test sample to evaluate, I would build the final model using all of the training data provided!

#### Question 7 ####

# If the Brewers had the option to purchase the 30th pick in the 2020 MLB Draft, how much would you recommend they pay to do so? Please explain your process for determining the value.

I would recommend that the Brewers pay $11 million for the 30th overall pick in the 2020 MLB Draft. I came to this number through research and articles published on Fangraphs and Baseball Prospectus. They viewed the 30th overall pick at roughly $10 million, Fangraphs was slightly higher than that Baseball Prospectus slightly lower, in value to the team. I believe the Brewers should be willing to pay above this pick valuation because the opportunity to have another first round pick, and its slot bonus, is incredibly valuable to a smaller market club, like the Brewers. The additional first round pick is important as it provides the Brewers with an additional player to develop and add to its competitive core/farm system who could potentially add roughly $10 million in surplus value to the organization.

Most importantly though is the value of the additional slot bonus is almost immeasurable; this would allow the Brewers to select players in other draft spots that could require above slot bonuses to sign and gain a competitive advantage over opponents that aren't able to "reach" on players due to a tighter signing bonus budget and concerns over player signability.

Mainly due to the increased signing bonus pool and the additional draft pick, I recommend that the Brewers pay $11 million for the 30th overall pick in the 2020 MLB draft.

#### Question 8 ####

# In 500 words or fewer, please describe potential implications of moving the Brewers' AAA affiliate from Colorado Springs, CO to San Antonio, TX.

There are a few main implications of moving the Brewers' AAA affiliate from Colorado Springs, CO to San Antonio, TX, most of them are beneficial to the Brewers and their AAA players! The main benefits of moving the affiliate are seen by the players.

Inner divisional travel is much easier on players. With the affiliate being placed in San Antonio the team will only need to travel Oklahoma City, OK, Round Rock, TX, and Wichita, KS, which are all closer than the previous in division cities. The farthest in division road trip is to Wichita, KS but is closer or equivalent in distance to all in division road trips from Colorado Springs. The shorter travel will help with player recovery and comfort as they spend less time on the road and more time sleeping and preparing for the next series. The new division created by playing in San Antonio instead of Colorado Springs also helps with player preparation and recovery because all divisional games will take place in the same time zone. Players will not have to travel across time zones therefore they will not struggle with sleep and body cycle adjustments before games. Travel will be a huge benefit and implication of moving the AAA affiliate to San Antonio.

Moving the affiliate out of Colorado Springs also will help player performance and evaluation. Pitchers will see the biggest improvement in on field results by leaving the high elevation of Colorado Springs and they will also have better results when refining their off-speed pitches. By moving the affiliate to San

Antonio pitchers will generate more consistent depth and break on their off-speed pitches because of the elevation change from 6,035 feet above sea level to 650 feet above sea level. The change will also help the Brewers evaluate pitchers because of the similar conditions to pitching at most MLB ballparks. This will allow the Brewers to better project their pitchers and create development plans to make them the most competitive at the next level. Moving the affiliate to San Antonio from Colorado Springs will greatly help with the Brewers' player development and evaluation.

By moving the Brewers' AAA affiliate from Colorado Springs, CO to San Antonio, TX will greatly help the players and make them happier. The improvement in travel conditions is a positive implication for players and the improved player evaluation and development implication is vital to an organization and continuing success at the major league level! Moving the affiliate to San Antonio, TX will greatly benefit the Milwaukee Brewers organization!