STAT 506: Midterm Exam
Name:

Please turn in the exam to GitHub and include the R Markdown code and a PDF or Word file with output. Please verify that all of the code has compiled and the graphics look like you think they should on your files, you are welcome to upload image files directly if they look distorted in the output file.

While the exam is open book and you can use any resources from class or freely available on the internet, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including group mates or class members.** All resources, including websites, should be acknowledged.

For full credit, include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands.

Short Answer Questions

For questions in this section, keep your answers concise. You are welcome to use a combination of prose, math, and pseudocode, but your responses should be well thought out and defended.

1. (4 points)

Describe statistical significance, in your own words.

Statistical significance is the concept that when assuming the null hypothesis is true, the data support a rejection of the null in favor of the alternative hypothesis given an amount of acceptable possibility of rejecting the null when the null is true.

2. (4 points)

How can the standard deviation of the data be used to characterize the uncertainty in a point estimate?

The standard deviation refers to the amount of variability that exists in the overall population. When making a point estimate of the mean, the standard deviation can be used to estimate the standard error, which describes the amount of variability that exists in the point estimate. The standard error for the mean is $\frac{\sigma}{\sqrt{n}}$*, and is used when creating confidence intervals.*

3. (4 points)

Convince a collaborator, say a scientist modeling butterfly movement, that using Bayesian analysis is a defensible approach.

A Bayesian analysis would be defensible for modeling butterfly movement because we have a lot of information that has been collected previously before this study that we could incorporate by using a prior distribution. It uses simulation-based methods that can be more robust to estimate parameters instead of exact methods that make assumptions that may or may not be satisfied.

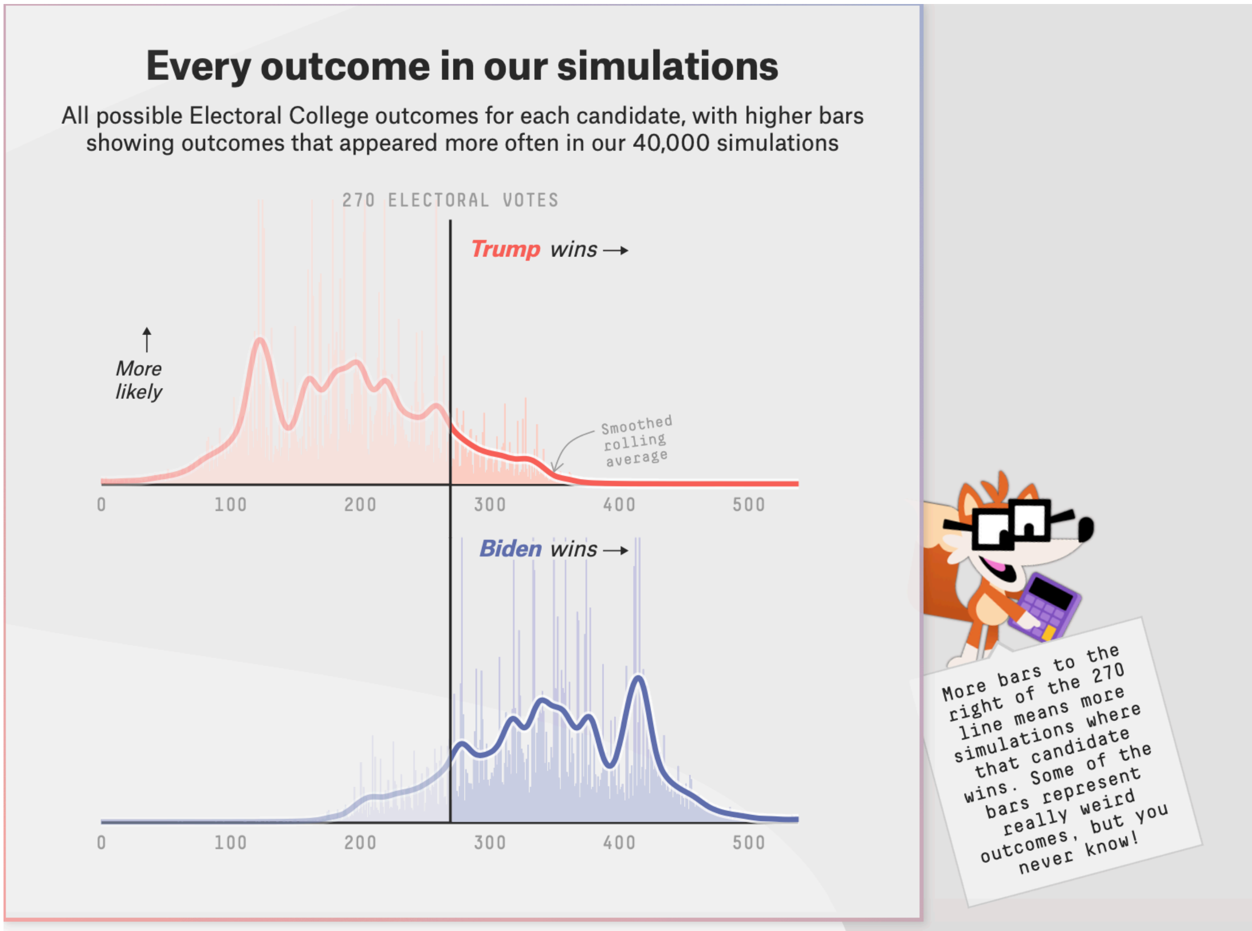
4. (4 points)

Convince a collaborator, say a scientist modeling butterfly movement, that a classical analysis is a defensible approach.

A classical analysis for modeling butterfly would be a defensible approach because it is using the data that we find in this study to estimate parameters using exact methods, which have a known amount of bias and variance associated with them.

5. (4 points)

Consider the following image, obtained from fivethirtyeight.com. Describe what the bars and curves on the figure represent *and* how the figure can be interpreted. You can assume your audience would be a STAT 216 student with an introductory knowledge of statistics.



.

Additional details about the figure and methods behind it are available at: https://projects.fivethirtyeight.com/2020-election-forecast/.

The height of the bars is the proportion of times that each candidate wins "x" number of electoral college votes when simulated 40,000 times, so the most likely outcomes are the tallest bars and the least likely are the shortest bars. The curve is an smoothed version of the bars to estimate the probability of each number of electoral college votes. The figure could be interpreted by looking at the proportion of simulations that are greater than or equal to 270 electoral college being the probability that each candidate wins enough of the electoral college to win the election. Since the majority of the probable number of votes that Biden gets is more than 270, Biden is expected to win.

Simulation Question

5. (10 points)

This question is focused on understanding an interaction with categorical variables. When answering this question, please include all code in the text.

a. Write out the mathematical model for a two-way Anova model with an interaction. (There should be at least 3 levels for one of the categorical variables, but the other can have two levels).

$\hat{Y} = \beta_0 + \beta_1 F_2 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 F_2 * G_2 + \beta_5 F_2 * G_3$

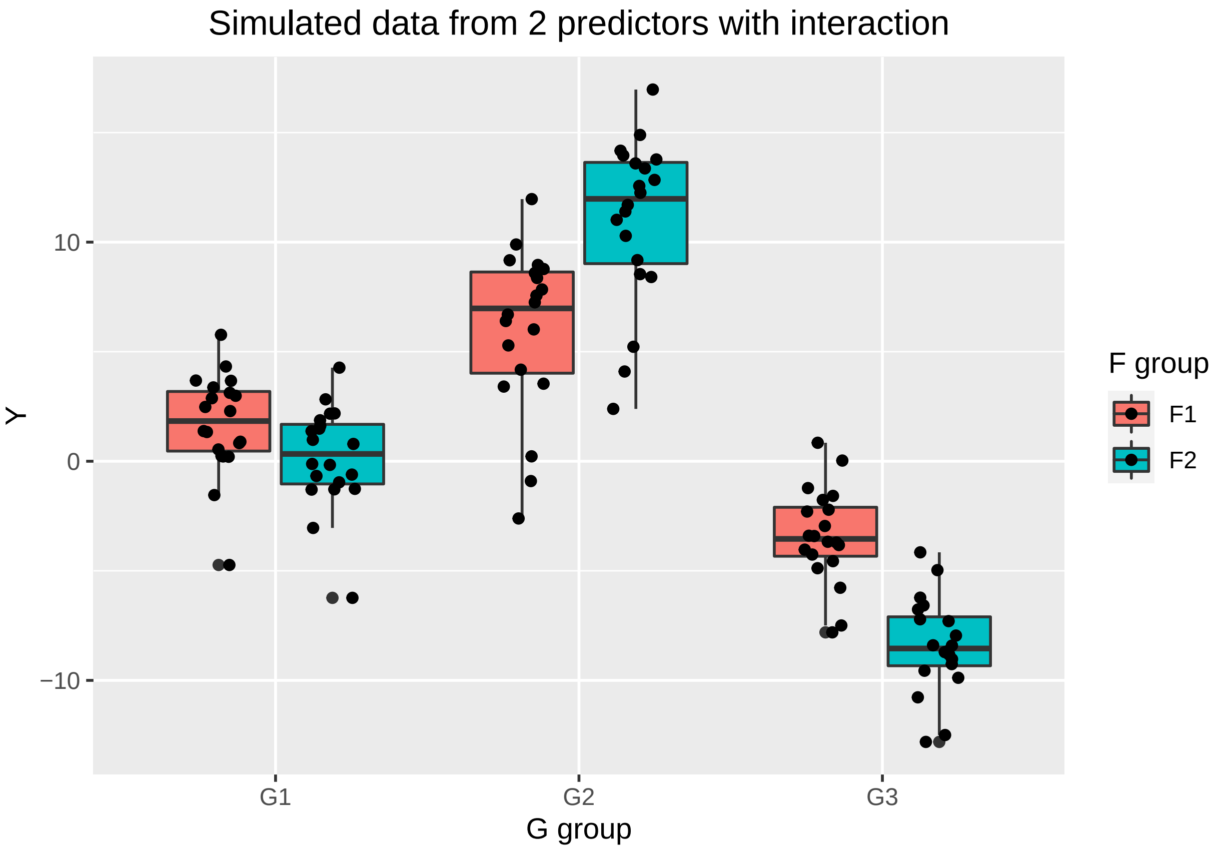
b. Simulate data from this model.

```r
set.seed(10202020)
beta0 = 1
beta1 = -1.5
beta2 = 5
beta3 = -5
beta4 = 10
beta5 = -10
SD = 3

Fgroup = rep(c("F1","F2"), each = 60)
Ggroup = rep(c("G1", "G2", "G3"), each = 20, times = 2)
err = rnorm(n = 60, mean = 0, sd = SD)

Y = beta0 +
  beta1*(Fgroup == 'F2' & Ggroup == 'G1') +
  beta2*(Fgroup == 'F1' & Ggroup == 'G2') +
  beta3*(Fgroup == 'F1' & Ggroup == 'G3') +
  beta4*(Fgroup == 'F2' & Ggroup == 'G2') +
  beta5*(Fgroup == 'F2' & Ggroup == 'G3') + err

df = data.frame(Y, Fgroup, Ggroup)
```

I don't know if I messed this up, but I was trying to create the β values with known quantities, then hard code the interactions through the model specification.

c. Create a data visualization that clearly depicts the interaction. Make sure to include appropriate titles, labels, and captions. Annotation may also be useful with this figure.

```r
ggplot(data = df, mapping = aes(x = Ggroup, y = Y, fill = Fgroup)) +
  geom_boxplot() +
  geom_jitter(position = position_jitterdodge(jitter.width = 0.15)) +
  labs(x = "G group",
       y = "Y",
       fill = "F group",
       title = "Simulated data from 2 predictors with interaction") +
  theme(plot.title = element_text(hjust = 0.5))
```



```
# https://stackoverflow.com/questions/46982172/split-geom-point-points-along-x-axis-by-group
# https://aosmith.rbind.io/2018/01/09/simulate-simulate-part1/
```

Since the median changes more for F2 (blue) compared to F1 (red) when going from G1 to G2 to G3, there is evidence of an interaction effect between F group and G group.

d. Fit the model and interpret the model coefficients.

```
model = stan_glm(data = df, Y ~ Fgroup * Ggroup, refresh = 0)
print(model)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      Y ~ Fgroup * Ggroup
##  observations: 120
##  predictors:   6
## ------
##                     Median MAD_SD
## (Intercept)          1.7    0.7  
## FgroupF2            -1.5    0.9  
## GgroupG2             4.3    0.9  
## GgroupG3            -5.1    0.9  
## FgroupF2:GgroupG2  6.5    1.4  
## FgroupF2:GgroupG3 -3.5    1.4  
## 
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 2.9    0.2
```

```
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
# True values
# beta0 = 1
# beta1 = -1.5
# beta2 = 5
# beta3 = -5
# beta4 = 10
# beta5 = -10
```

The model estimates for the β_0 through β_3 are fairly accurate, within 1 SD of the true value. For β_4 and β_5, the model estimates for the values are way off. The only thing that I can think of is that maybe there was some automatic centering going on or something for those estimates, since they add up to the correct value, or that I messed up the simulation process somehow. The estimate for sigma is very close to the standard deviation of the errors sampled from a normal distribution, which I thought was interesting.

Reading and Critique

6. (10 points)

Read the article titled, "Survey data on students' online shopping behaviour: A focus on selected university students in Indonesia" (link: https://www.sciencedirect.com/science/article/pii/S2352340919314295)

a. Comment on the experimental design and how the results from the study can be generalized.

The authors sampled 393 students by randomly selecting from the student ID numbers of students from an Indonesian university of 20,000 students, mostly in their 5th and 7th semesters. Those students filled out a survey of Likert scale 1-5 about perceptions of online shopping behavior if they had online shopped before.

The sampling scheme of randomly sampling student IDs, assuming it was done randomly and fairly, should have been reasonable. The sample should be able to generalize to all students at the university attending in 2020. There is no ability to make any causal inference because there was no way to randomly assign students to treatment groups. The authors probably could have considered a weighted sampling scheme or a stratified sampling scheme to guarantee getting a representative sample.

b. To the best of your ability, write out the mathematical notation that corresponds to the regression model displayed in Tables 7, 8, and 9.

$$\widehat{ShoppingBehavior} = \beta_0 + \beta_1 POR + \beta_2 TAS + \beta_3 QOW + \beta_4 EJY + \beta_5 SIF + \beta_6 OAD + error$$

c. Critique the following section.

> The hypothesis to be tested is as follow: Ho: There are no variables influencing online shopping behaviour.
>
> We see that the ANOVA produces P-value of the regression = 0.000, which is less than 0.05 significant level. This leads to the rejection of the null hypothesis, meaning that at least one of the predictors significantly influences the purchasing behaviour. The R-square is 47.26%, meaning that the predictors have an effect of 47.26% on onine shopping behaviour.
>
> The coefficients in Table 8 show the individual effect of each variable. We see that the P-values of POR, EJY, SIF and OAD are less than 0.05 significant level. This means that the purchasing behaviour is significantly infuenced by the perception of risk (POR), enjoyment (EJY), social influence (SIF) and online advertisment (OAD). Meamwhile, two other variables, i.e. trust and security (TAS) and quality of website (QOW), did not significantly influence the online shopping behaviour (see Table 9).

I think the authors should have re-read their paper to find grammatical errors ("is as follow", "the ANOVA produced P-value of the regression = 0.000", "onine shopping behavior", etc).

Their hypothesis that they stated is not the one that they are actually testing. They said that the H_0 is that there are no variables that influence online shopping behavior, but what the actual F-test is testing is 'does the regression model explain the variation in the data better than a model using only the mean', but it doesn't test the actual variables individually like the authors suggest.

The authors suggest that a P-value could be equal to 0.000. I would probably instead say 'less than 0.001'.

Their interpretation of R^2 is not correct… they say that the predictors "have an effect of 47.3% on online shopping behavior", but it's not clear what the authors mean by 'an effect'. A clearer interpretation would be "47.3% of the variation in the online shopping behavior is explained by the model.

When interpreting the p-values, I would have included what the p-values were for each variable, especially for non-significant predictors. There is a difference between p = 0.05001 and p = 0.74 for what I would consider to be important. I know that information is in the table but I still think it's useful to include.

d. Based on the analysis (and assuming all assumptions are satisfied). Write a summary statement of the results in Table 9.

There are 4 statistically significant predictors of online shopping behavior: perception of risk (p = 0.022), enjoyment (p = 0.026), social influence (p = 0.035), and online advertisement (p= 0.015). Trust and security (p = 0.744) and quality of website (p = 0.179) were not significant. Of the statistically significant predictors, enjoyment has the largest effect (coef = 0.314 ± 0.138), followed by SIF (coef = 0.264 ± 0.123), perception of risk (coef = -0.254 ± 0.109), and online advertisement (coef = 0.170 ± 0.068).

Data Analysis

7. (10 points)

Using a candy dataset (https://math.montana.edu/ahoegh/teaching/stat446/candy-data.csv), define and fit a regression model to understand the relationship between `winpercent` and `pricepercent, chocolate,` and `caramel`. More insight into the data is available at https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/.

a. Write out the model and define all of the coefficients.

$$WinPercent = \beta_0 + \beta_1 PricePercent + \beta_2 Chocolate + \beta_3 Caramel + \beta_4 Chocolate * Caramel + error$$

β_0 is the win percentage for a candy that is free, with no caramel or chocolate, β_1 is the change in win percentage for a 1 percentile increase in price, β_2 is the change in win percentage for a candy that contains chocolate and not caramel, holding price constant, β_3 is the change in win percentage for a candy that contains caramel and not chocolate, holding price constant, β_4 is the change in win percentage from the addition of chocolate and caramel due to an interaction effect between the two.

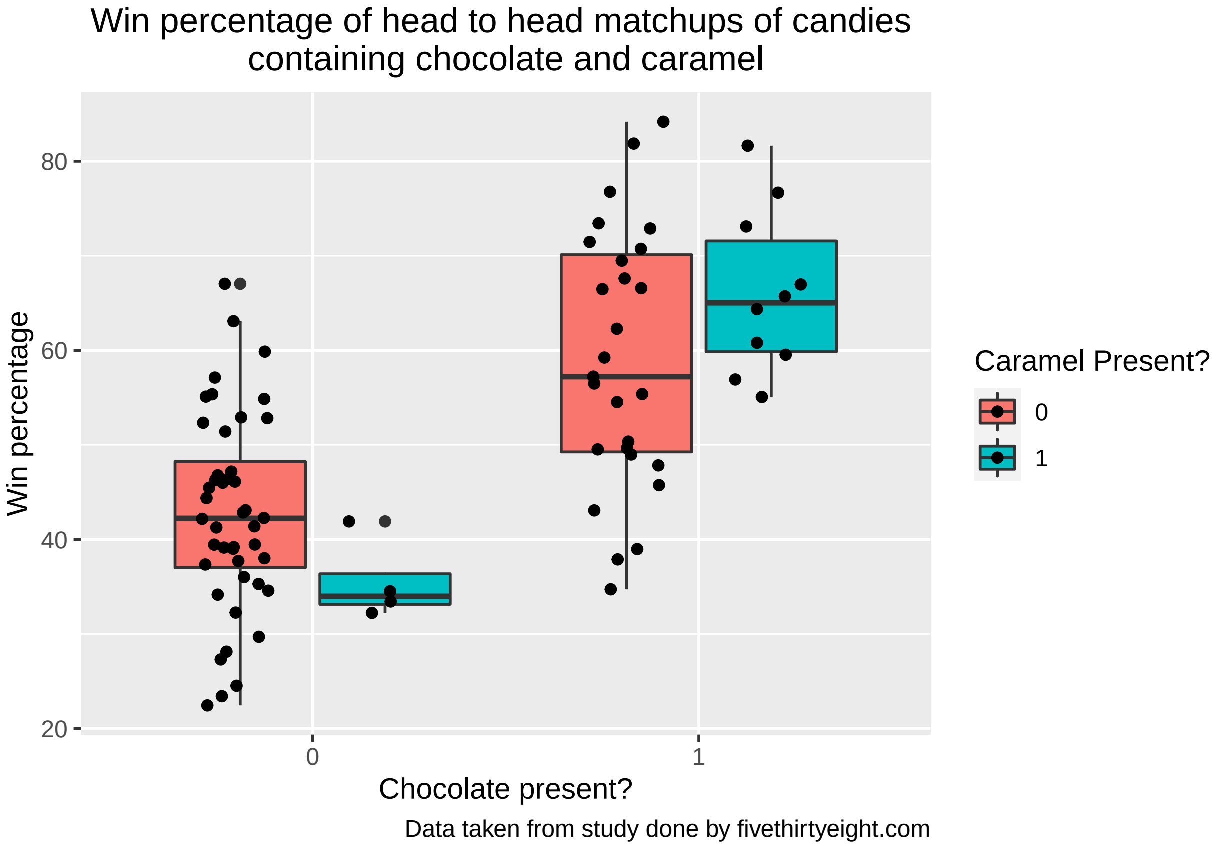
b. Fit the model with software of your choice and print the results.

```
candy = read_csv("https://math.montana.edu/ahoegh/teaching/stat446/candy-data.csv")
```

```
## Parsed with column specification:
## cols(
##   competitorname = col_character(),
##   chocolate = col_double(),
##   fruity = col_double(),
##   caramel = col_double(),
##   peanutyalmondy = col_double(),
##   nougat = col_double(),
##   crispedricewafer = col_double(),
##   hard = col_double(),
##   bar = col_double(),
##   pluribus = col_double(),
##   sugarpercent = col_double(),
##   pricepercent = col_double(),
##   winpercent = col_double()
## )
```

```
cols = colnames(candy[,2:10])
candy[,2:10] = lapply(candy[,cols], as.factor)
```

```
ggplot(data = candy, mapping = aes(y = winpercent, x = chocolate, fill = caramel)) +
  geom_boxplot() +
  geom_jitter(position = position_jitterdodge()) +
  labs(x = "Chocolate present?",
       y = "Win percentage",
       fill = "Caramel Present?",
       title = "Win percentage of head to head matchups of candies \ncontaining chocolate and caramel",
       caption = "Data taken from study done by fivethirtyeight.com") +
  theme(plot.title = element_text(hjust = 0.5))
```



There is evidence that there is an interaction effect between chocolate and caramel, since the median win percent increases more for candies with caramel compared to those without caramel when chocolate changes from absent to present.

```
set.seed(10252020)
model = stan_glm(data = candy, formula = winpercent ~ pricepercent + (chocolate * caramel), refresh= 0)
print(model)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      winpercent ~ pricepercent + (chocolate * caramel)
##  observations: 85
##  predictors:   5
## ------
##                       Median MAD_SD
## (Intercept)           42.4    2.5
## pricepercent           0.9    5.0
## chocolate1            16.0    3.1
## caramel1              -7.3    6.0
## chocolate1:caramel1 14.3    7.2
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 11.4    0.9
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

When chocolate and caramel being present is held constant, a one unit increase in price (which would be going from the cheapest to the most expensive candy) only contributes a 0.9% (± 5.0%) increase in win percentage of the candy, which makes sense because of the study design. Participants were asked to "Please select which fun-size Halloween treat you would most want to receive as a trick-or-treater", which implies that price shouldn't be a factor because both candies are free.

When price is held constant and caramel is not present, having chocolate in the candy is associated with a 16.0% (± 3.1%) increase in win percentage. When price is held constant and chocolate is not present, having caramel in the candy is associated with a 7.3% decrease *(± 6.0%) in win percentage. This makes sense, many popular candies have chocolate but not caramel, like all of Reese's products, M&M's, and Kit-kat bars, but not many candies have caramel but not chocolate and the ones that do are not very popular.*

When price is held constant and both caramel and chocolate are present, there is an additional increase in win percentage by 14.3% (± 7.2%) due to the combination of caramel and chocolate on a candy. This makes sense, because a lot of very popular candies contain both chocolate and caramel, like Snickers, Twix, and Milky Way.

c. Construct a contrast to compare the expected win percentage for a candy that has chocolate, caramel and the average price percent, with a candy that just has chocolate, no caramel, and the average price percent.

```r
# From Ch 10 notes
# Find contrasts
options(warn = 0)
dat = as.data.frame(model) %>%
  mutate(contrast = 0 + mean(pricepercent) - mean(pricepercent) + chocolate1 + caramel1 + chocolate1:ca
```

```
## Warning: Problem with `mutate()` input `contrast`.
## i numerical expression has 4000 elements: only the first used
## i Input `contrast` is `+...`.
```

```
## Warning in chocolate1:caramel1: numerical expression has 4000 elements: only the
## first used
```

```
## Warning: Problem with `mutate()` input `contrast`.
## i numerical expression has 4000 elements: only the first used
## i Input `contrast` is `+...`.
## Warning in chocolate1:caramel1: numerical expression has 4000 elements: only the
## first used

## Warning: Problem with `mutate()` input `contrast`.
## i longer object length is not a multiple of shorter object length
## i Input `contrast` is `+...`.

## Warning in 0 + mean(pricepercent) - mean(pricepercent) + chocolate1 + caramel1
## + : longer object length is not a multiple of shorter object length
```

```
dat %>% summarize(median_diff = median(contrast),
            lower_interval = quantile(contrast, probs = 0.025),
            upper_interval = quantile(contrast, probs = 0.975))
```

```
##   median_diff lower_interval upper_interval
## 1    11.61209      -5.683172        29.0333
```

```
# # Not sure what this is doing, wanted to make predictions for the two contrasts
# # Set up groups to compare
# df = data.frame(pricepercent = c(rep(mean(candy$pricepercent, 2))),
#                  chocolate = factor(c(1, 0)),
#                  caramel = factor(c(1, 0)))
# # Find median of the posterior predicted value for those two
# apply(posterior_predict(model, newdata = df), 2, median)
```

There is a predicted increase in win percentage of 11.6% (95% CI: -5.7%, 29.0%) for a candy that has chocolate and caramel compared to a candy with neither chocolate nor caramel when the price is the mean price for both candies.

d. Try to create the candy with the highest win percentage, then specify the levels of the predictors and create a predictive distribution for an individual type of candy with those features.

```
set.seed(10252020)
candy2 = candy %>% select(-competitorname)
model2 = stan_glm(data = candy2, winpercent ~ . + chocolate * caramel, refresh = 0)
print(model2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      winpercent ~ . + chocolate * caramel
##  observations: 85
##  predictors:   13
## ------
##                      Median MAD_SD
## (Intercept)          36.7    4.7  
## chocolate1           17.9    4.1  
## fruity1               8.0    4.1  
## caramel1             -3.7    6.3  
## peanutyalmondy1       9.6    3.7  
## nougat1              -1.0    5.7  
## crispedricewafer1     7.8    5.5  
## hard1                -6.1    3.5  
## bar1                 -0.2    4.9  
## pluribus1            -1.7    3.1  
## sugarpercent          9.2    4.7  
## pricepercent         -5.7    5.5  
## chocolate1:caramel1   8.9    7.7  
##
```

```
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 10.7    0.9  
## 
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

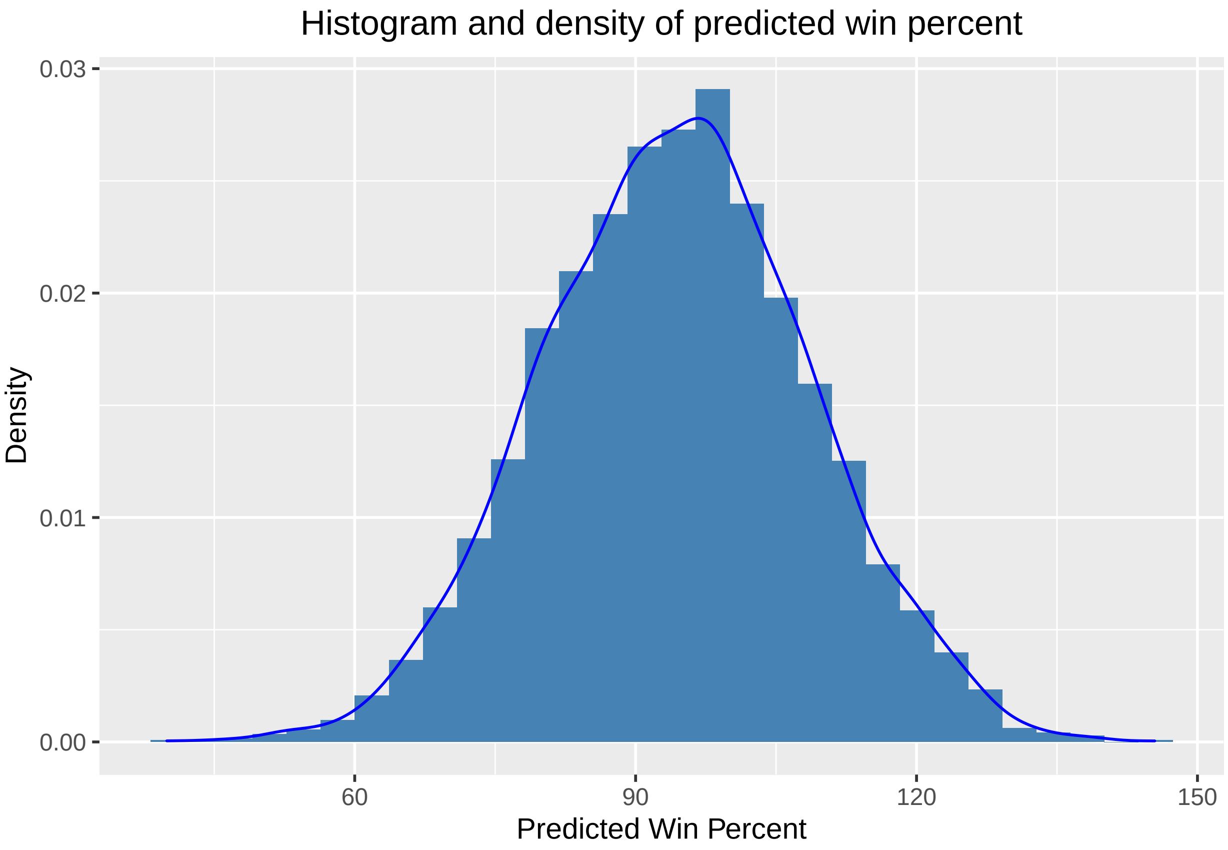
After fitting a model with every predictor, the optimal candy has chocolate, fruity flavor, caramel, peanuts/almonds, crisped rice wafer, the highest sugar percent possible, and the lowest possible price.

```r
df2 = data.frame(chocolate = as.factor(1),
                 fruity = as.factor(1),
                 caramel = as.factor(1),
                 peanutyalmondy = as.factor(1),
                 nougat = as.factor(0),
                 crispedricewafer = as.factor(1),
                 hard = as.factor(0),
                 bar = as.factor(0),
                 pluribus = as.factor(0),
                 sugarpercent = max(candy$sugarpercent),
                 pricepercent = min(candy$pricepercent))
pred = posterior_predict(model2, df2)
median(pred);quantile(pred, c(0.025, 0.975))
```

```
## [1] 94.36678

##      2.5%     97.5% 
##  66.41065 122.59755
```

```r
ggplot(mapping = aes(x = pred, y = ..density..)) +
  geom_histogram(fill = "steelblue") +
  geom_density(col = "blue") +
  labs(title = "Histogram and density of predicted win percent",
       x = "Predicted Win Percent",
       y = "Density") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Don't know how to automatically pick scale for object of type ppd/matrix/array. Defaulting to continu

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The perfect candy (with chocolate, fruity flavor, caramel, peanuts/almonds, crisped rice wafer, and sugar) sounds disgusting but would be predicted to win around 94.4% of the time (95% CI: 66.4%, 100%). I don't think that this super candy would actually win this often, because everyone knows that the Snickers is the best candy ever, and I think a candy with all those features would have too many flavors going on.