Exam # 2

Sunny Lee

2021-03-24

1) Categorical dependent variables. You want to test to see if there is a relationship between whether or not someone has blue eyes and whether or not they have blonde hair. You collect data and observe the results in the following table.

```
Bird<-cbind(c(39,90),c(86,273))
rownames(Bird)<-c("Blond Hair", "Not Blond Hair")
colnames(Bird)<-c("Blue eyes","Not Blue eyes")
Bird
```

```
##                Blue eyes Not Blue eyes
## Blond Hair            39            86
## Not Blond Hair        90           273
```

```
sum(Bird)
```

```
## [1] 488
```

a) If the two variables are independent, what would you estimate is the probability of observing exactly 39 blue eyed blonde haired people in the group of this size?

From the table above, we can find the probability of observing exactly 39 blue eyed blonde haired people by: (359C86)(129C39)/(488C125) =

```
choose(359, 86) * choose(129, 39) / choose(488, 125)
```

```
## [1] 0.03484396
```

```
dhyper(39, 129, 359, 125)*100
```

```
## [1] 3.484396
```

And so we would expect a 3.48 percent chance of observing exactly 39 blue eyed blonde haired people.

b) If the two variables are independent, how many blue eyed blonde haired people would you expect in a group of this size?

We can calculate our expected value by taking the column total and multiplying it by the row total over the over all total:

```
129 * (125/488)
```

```
## [1] 33.04303
```

and so our expected value for a blue eyed blonde haired person is 33.04 in our group.

c) What is the probability of observing a number of blonde haired blue eyed people which is at least as far from expected as we did?

```
sum(dhyper(39:125, 129, 359, 125))
```

```
## [1] 0.1005819
```

We would expect a probability of 10.05% chance of observing at least 39 blonde haired blue eyed people.

d) Create the expected count table and calculate χ^2 test statistic.

```
expectedBird<-cbind(c(33.04303,95.95697),c(91.95697,267.043))
rownames(expectedBird)<-c("Blond Hair", "Not Blond Hair")
colnames(expectedBird)<-c("Blue eyes","Not Blue eyes")
expectedBird
```

```
##                  Blue eyes Not Blue eyes
## Blond Hair        33.04303      91.95697
## Not Blond Hair    95.95697     267.04300
```

```
sum(expectedBird)
```

```
## [1] 488
```

```
chisq <- 0
for(i in 1:2){
  for (j in 1:2){
    chisq <- chisq + ((Bird[i, j] - expectedBird[i, j])^2 / expectedBird[i, j])
  }
}

chisq
```

```
## [1] 1.962501
```

So, our Chi-Squared test statistic is 1.962501.

e) Use a χ^2 test to either reject or accept the null hypothesis that these to variable are independent. Explain your results with a p-value.

```
#checking to see if proportion of blue eyes is same in blonde and non blonde people
qchisq(.975, df = 1)
```

```
## [1] 5.023886
```

```
qchisq(.025, df = 1)
```

```
## [1] 0.0009820691
```

```
pchisq(1.962501, df = 1, lower.tail = FALSE)
```

```
## [1] 0.1612461
```

We see from our 97.5th and .025th percentiles of a Chi square distribution that our test statistic is within the two values, and thus we fail to reject the null hypothesis. We can also see from our p value that we fail to reject the null hypothesis since it is greater than $\alpha = .05$. Thus, we conclude that the proportion of blue eyed people is the same in blonde and not blonde groups.

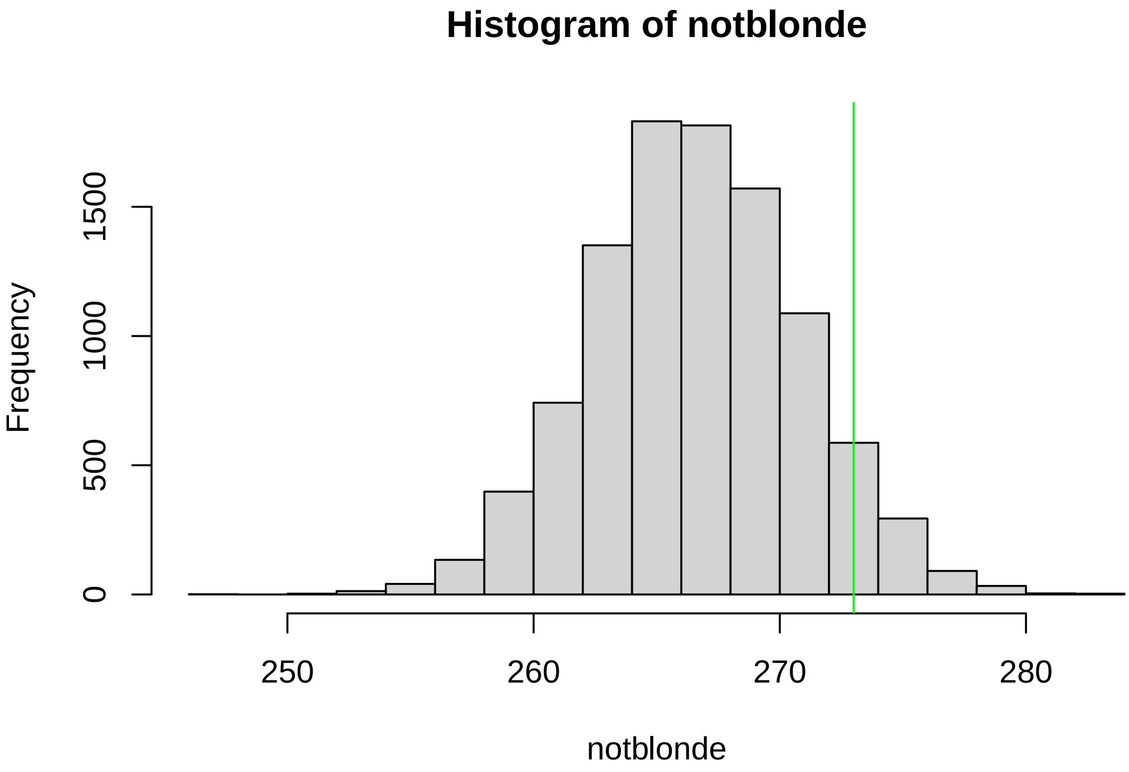
f) Create a simulation study to test the one-sided hypothesis that not blond hair are more likely not a blue eye. Cleary state the null and alternative hypothesis, give the p-value, and provide a conclusion. Run a simulation for 10,000.

If we want to check if the not blonde hair are more likely to not have blue eyes, our null hypothesis must be that the proportion of not blue eyes in the not blonde group is equal to the proportion of not blue eyes in the blonde group: $H_0 : P_{NB} = P_B$. Thus, our alternative hypothesis must be that the not blue eyes in the not blonde group is higher than in the blonde group: $H_a : P_{NB} > P_B$.

```
haircolor<-c(rep("blue",129),rep("notblue",359))
exp1<-sample(haircolor,363,replace=FALSE)
sum(exp1=="notblue")
```

```
## [1] 266
```

```
p=10000
notblonde<-rep(0,p)
for(i in 1:p)
{
  temp<-sample(haircolor,363,replace=FALSE)
  notblonde[i]<-sum(temp=="notblue")
}
hist(notblonde)
abline(v=273,col="green")
```



```
p_value<-mean(notblonde>=273)
p_value
```

```
## [1] 0.1012
```

By looking at the p value = .1011 we obtain, we see that it is greater than .05, and thus we fail to reject the null hypothesis. Since we fail to reject the null hypothesis, we conclude that the proportions of the not blue eyed people are the same in the blonde and not blonde groups.

g) Use Fisher's exact test to test the one sided hypothesis that not blond hair are more likely to be not a blue eye. Compare the p-value with the p-value of the simulation.

```
sum(dhyper(273:363, 359, 129, 363))
```

```
## [1] 0.1005819
```

Using Fisher's exact test, we get an exact p value of 0.1005819. We see that this value is quite close to our simulated value of 0.1011.

2) The next question should be answered using the carseats data set.

```
library(ISLR)
data(Carseats)
attach(Carseats)
```

a) Fit a multiple regression model to predict Sales using Price, Urban, and US. Give summary of the model and interpret R^2.

```
model <- lm(Sales~Price+Urban+US)
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

From the summary of the model, we see a low p value for all coefficients except for Urban. Thus, in our final model, we would do well to remove Urban from our explanatory variables. The F-statistic for this model is also quite low, so we can conclude that our model is significant in predicting Sales. We also have a low R^2 value, meaning our model is not very great at predicting Sales. Thus, while our model is significant, it is not yet a very good model for predicting Sales.

b) Provide an interpretation of each coefficient in the model. Be careful some of the variables in the model are qualitative.

For the intercept, it is the initial sales when all other variables are zero. For Urban and US, these coefficients are the initial change when the coefficient for Price is zero. The Price coefficient is how much Sales changes with Price if Urban and US are zero.

c) Write out the model in equation form, being careful to handle the qualitative variables properly.

$Sales = 13.043469 - 0.054459(Price) - 0.021916(Urban) + 1.200573(US)$ where $Urban, US \in \{0, 1\}$.

d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We can reject the null hypothesis for Intercept, Price, and US, however not for Urban.

e) Now, fit a new model with interaction term to predict Sales using $price * Urban$ and $Price * US$. Write out the model in equation form and interpret the coefficients in the context of the problem.

```
model1 <- lm(Sales~Price*Urban+Price*US)
summary(model1)
```

```
##
## Call:
## lm(formula = Sales ~ Price * Urban + Price * US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8978 -1.6033 -0.0639  1.5992  7.1211
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.855670   1.348612  10.274  < 2e-16 ***
## Price          -0.061555   0.011564  -5.323 1.72e-07 ***
## UrbanYes       -1.265181   1.370412  -0.923    0.356
## USYes           1.337245   1.255007   1.066    0.287
## Price:UrbanYes  0.010844   0.011709   0.926    0.355
## Price:USYes    -0.001232   0.010670  -0.116    0.908
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.476 on 394 degrees of freedom
## Multiple R-squared:  0.2409, Adjusted R-squared:  0.2313
## F-statistic: 25.01 on 5 and 394 DF,  p-value: < 2.2e-16
```

$Sales = 13.855670 - 0.061555(Price) - 1.265181(Urban) + 1.337245(US) + 0.010844(Price * Urban) - 0.001232(Price * US)$. The intercept, Price, Urban, and US are the same as above. The Price * Urban coefficient changes how Sales is affected by Price if Urban is 1. The Price * US coefficient changes how Sales is affected by Price if US is 1.

f) Conduct extra sum of squares test to determine if interaction terms are important to the model. Clearly define the null and alternative hypothesis interms of $\beta_i's$ only.

For the Extra Sum of Squares test, our null hypothesis is that all extra coefficients are equal to zero: $H_0 : \beta_5 = \beta_6 = 0$, and so our alternative hypothesis is that these coefficients are not equal to zero: $H_a : \beta_5 \neq \beta_6 \neq 0$.

```
anova(model, model1)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price * Urban + Price * US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 2420.8
## 2    394 2415.5  2    5.2935 0.4317 0.6497
```

From the p value of our extra sum of squares test, we fail to reject the null hypothesis. Thus, we conclude that our new coefficients are not important to model our data. We can also see this from our summary of model1, where none of the interaction terms have low p values.

3) Assume we want to use three cooking times for popping the popcorn of Fastco brand. The popcorn is cooked in a microwave at 105, 120, and 135 seconds.

```
Popcorn<-cbind(c(82.60,81.07,77.70),c(87.50,90.80,72.54),c(75.30,84.06,65.98),c(80.20,71.50,72.56),c(73
rownames(Popcorn)<-c("105 Sec.","120 Sec.","135 Sec.")
Popcorn
```

```
##           [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
## 105 Sec. 82.60 87.50 75.30 80.20 73.80 77.60 88.50 83.56
## 120 Sec. 81.07 90.80 84.06 71.50 81.34 92.60 74.30 83.36
## 135 Sec. 77.70 72.54 65.98 72.56 74.51 80.83 88.97 70.44
```

a) Create a histogram of the 10,000 randomization/permutation test for the difference in means between 105 and 120 seconds. State the hypothesis test and calculate the p_value and include the R-code.

For the randomization/permutation test, our null hypothesis is our two cooking means are equal to each other $H_0 : \mu_105 = \mu_120$ and our alternative hypothesis is that the average cooking mean is greater in 120 sec. vs 105 sec. $H_a : \mu_120 > \mu_105$.

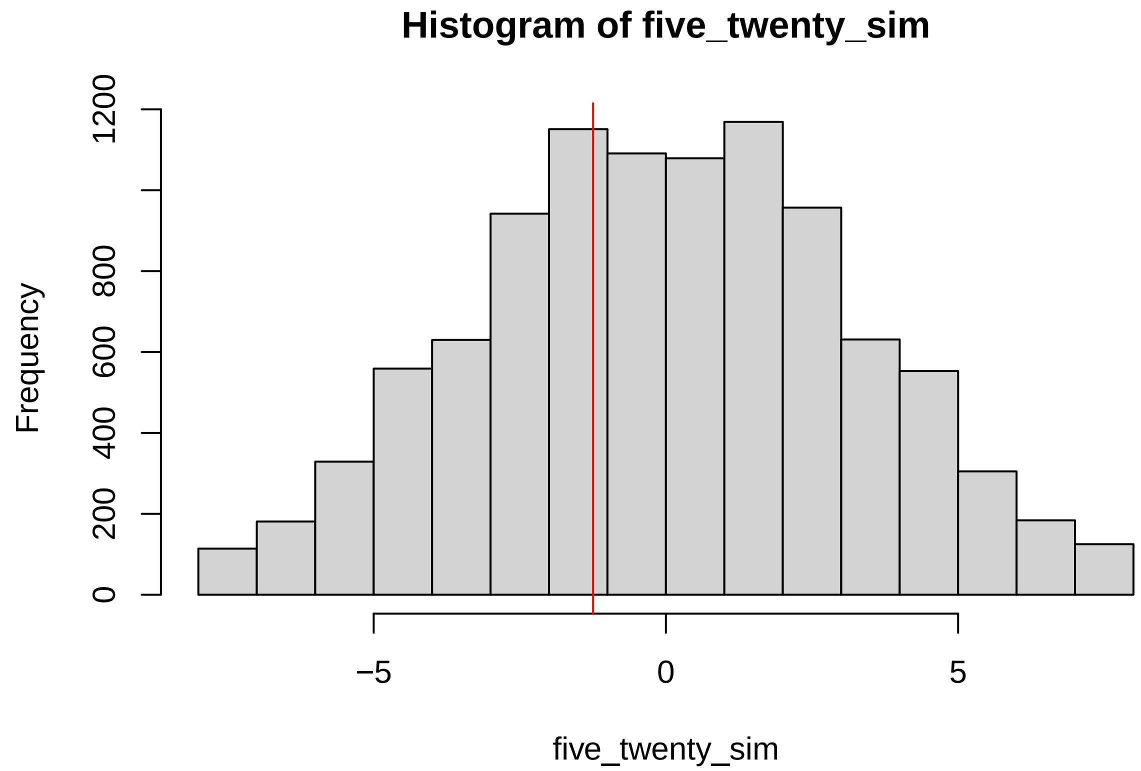
```
five_twenty <- Popcorn[1, ] - Popcorn[2, ]
obs <- mean(five_twenty)
obs
```

```
## [1] -1.24625
```

```
multiplier <- sample(c(1, -1), length(five_twenty), replace = TRUE)

mean(five_twenty * multiplier)
```

```
## [1] 1.47875
```

```
p <- 10000
five_twenty_sim <- rep(0, p)

for (i in 1:p){
  tempMult <- sample(c(1, -1), length(five_twenty), replace = TRUE)
  five_twenty_sim[i] <- mean(tempMult * five_twenty)
}

hist(five_twenty_sim)
abline(v = obs, col = "red")
```



```
p_value <- mean(five_twenty_sim > obs)
p_value
```

```
## [1] 0.644
```

Looking at the histogram, we can see that our observed difference is quite close to the mean of our simulated difference. Looking at the p value, we support this observation and, since our p value is greater than .05, we fail to reject the null hypothesis. Thus, we conclude cooking the popcorn for 105 and 120 seconds seems to cook the popcorn the same way.

b) Calculate the Wilcoxon Rank sum test statistic W_{105sec} to compare if the distribution of 105 second is the same as 120 second.

```
wilcox.test(Popcorn[1, ], Popcorn[2, ])
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Popcorn[1, ] and Popcorn[2, ]
## W = 28, p-value = 0.7209
## alternative hypothesis: true location shift is not equal to 0
```

From the wilcox.test, we get a test statistic of W = 28.

c) Use a software to conduct the Wilcoxon rank sum test to determine if the distribution of 105 Sec is the different from the 120 sec.

The null hypothesis for the Wilcoxon rank sum test is that the groups have the same distribution and our alternative hypothesis is that our two groups have different distributions. Looking at the p value, it is higher than .05, and thus we fail to reject the null hypothesis. Since we fail to reject the null hypothesis, we conclude that the two groups have the same distribution.

d) Use a software to run the Kruskal-Wallis test to determine if the distribution of Popcorn differs by time.

The null hypothesis for the Kruskal-Wallis test is all of our distributions are similar and the alternative is that at least one of the distributions are different.

```
kruskal.test(Popcorn[1, ], Popcorn[2, ], Popcorn[3, ])
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Popcorn[1, ] and Popcorn[2, ]
## Kruskal-Wallis chi-squared = 7, df = 7, p-value = 0.4289
```

After running the Kruskal-Wallis test, we observe a p value fo .4289, which is greater than .05. This means we fail to reject the null hypothesis. Since we fail to reject the null hypothesis, we conclude that the three distributions are the same.

4) Assume we use a least square regression to estimate parameter coefficients and found it to be $\hat{\beta}_0 = 2.024$ and $\hat{\beta}_1 = -4.024$. The response variable is $y = 3, 2, 1, 9, 12$ and explanatory variable is $x = 2, 3, 3.5, 4.3, 6$. Note: You can't use R for this data as the coefficients are not the true predicted values from the regression fit.

a) Calculate the sum of square error.

```
y = c(3, 2, 1, 9, 12)
x = c(2, 3, 3.5, 4.3, 6)

beta1 <- -4.024
beta0 <- 2.024

yhat <- beta1*x + beta0
yhat
```

```
## [1]  -6.0240 -10.0480 -12.0600 -15.2792 -22.1200
```

```
SSE = sum((y - yhat)^2)
SSE
```

```
## [1] 2150.804
```

b) Calculate the sum of square regression.

```
SSR = sum((mean(y) - yhat)^2)
SSR
```

```
## [1] 1858.98
```

c) Calculate the total sum of squares.

```
SST = SSR + SSE
SST
```

```
## [1] 4009.784
```