CDA Midterm

Joshua Ingram

11/5/2019

1. Pedestrians, Car Accidents, and Race

NPR covered a story on traffic accidents resulting in fatalities of pedestrians. They interviewed various researches interested in this topic and the said they found a signficant difference in the proportions of people that stopped for black people vs white people. Within this study they conducted, they had a number of subjects, some white and some black, step into the street to see if a car would stop for them in Las Vegas. It turns out that people were more likely to stop for the white subjects than the black subjects.

Below is a table of the observed data:

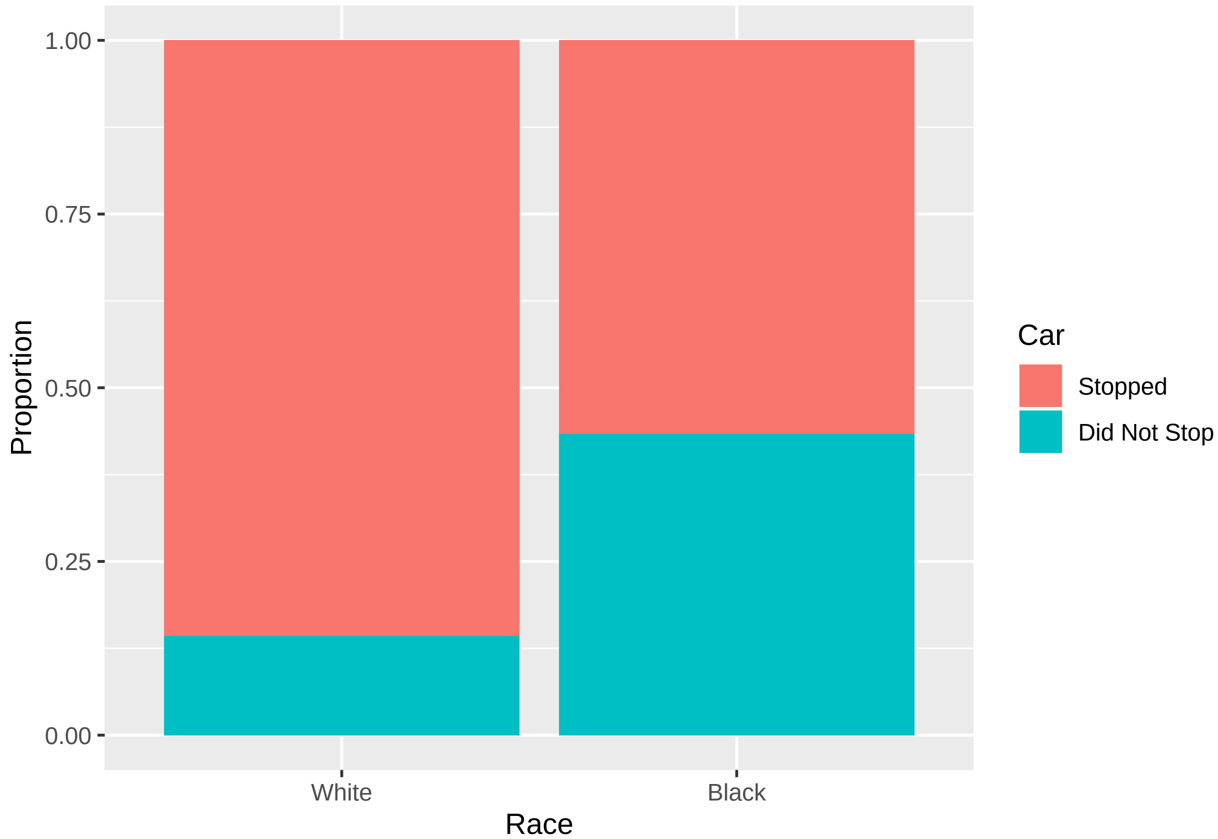
```
pedes
```

```
##       Stopped Did Not Stop
## White      24            4
## Black      17           13
```

There are number of things that could be the cause of this difference, whether it be racial bias, location, or some other confoudning variable. With the data in our hands, let's see what we find.

Data Visualization

```
ggplot(data = pedes.frame.cond, aes(x=Race, y=Proportion, fill = Car)) + geom_bar(stat="identity")
```



As we can see, there is a noticeable difference in the proportion of cars that did not stop for black pedestrians to proportion cars that did not stop for white pedestrians.

Chi-squared Test for Independence

First, let's see if race and whether the car stopped are independent of each other.

Ho: Race and whether the car stops are independent Ha: Race and whether the car stops are not independent $\alpha = 0.05$

```
expected <- chisq.test(pedes)$expected
expected
```

```
##       Stopped Did Not Stop
## White 19.7931     8.206897
## Black 21.2069     8.793103
```

After finding the expected counts, we see that there are at least 5 expected counts in each cell. It is safe to use the Chi-Squared Test.

```
chisq.test(pedes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  pedes
## X-squared = 4.5792, df = 1, p-value = 0.03236
```

We received a X^2 test statistic of 4.58 that yielded us a p-value of 0.03236. With this p-value, we have significant evidence to reject the null hypothesis, that race and whether cars stop are independent, and evidence that suggests there is not independence between the two. Knowing this, let's take a look some relevent statistics that can shed more light.

Relevent Statistics

Difference in proportions and Relative Risk:

Difference in Proportions:

```r
#test for difference in proportions
prop.test(c(pedes[2,2],pedes[1,2]), c(sum(pedes[2,]),sum(pedes[1,])), conf.level = 0.95, correct = FALSE
```

```
## $statistic
## X-squared
##  5.897875
##
## $parameter
## df
##  1
##
## $p.value
## [1] 0.01515916
```

After testing to see if the proportion of cars that did not stop for black people were equal to that of white people, we received a p-value of 0.01516, giving us significant evidence to believe the proportions are not equal.

```r
# difference in proportions
diff.prop <- pedes.cond[2,2] - pedes.cond[1,2]
diff.prop
```

```
## [1] 0.2904762
```

Based on our sample, we estimate a difference in proportions of 0.29, where the proportion of cars that did not stop for blacks is 29 percentage points higher than that for whites.

```r
#confidence interval
diffscoreci(pedes[2,2], sum(pedes[2,]), pedes[1,2], sum(pedes[1,]), conf.level = 0.95)
```

```
##
##
##
## data:
##
## 95 percent confidence interval:
##  0.05663533 0.49752447
```

We are also 95% confidence that the true proportion of cars that did not stop for black people is between 5.67 and 49.75 percentage points higher than the proportion of cars that did not stop for white people in Las Vegas.

Relative Risk:

```
#ratio
relative.risk <- pedes.cond[2,2] / pedes.cond[1,2]
relative.risk
```

```
## [1] 3.033333
```

```
# confidence interval
riskscoreci(pedes[2,2], sum(pedes[2,]), pedes[1,2], sum(pedes[1,]), conf.level = 0.95)
```

```
##
##
##
## data:
##
## 95 percent confidence interval:
##  1.212177 8.148850
```

We are 95% confident that a car is between 21.21% and 8 times more likely to not stop for a black pedestrian over a white pedestrian. Our point estimate, based on our observations, is that cars are 3 times more likely to not stop for black pedestrians than white pedestrians in Las Vegas.

Statistical Analysis

After our tests, we found through the Chi-Squared Test for Independence that race and whether a car stops are not independent of each other. With this in mind, we moved forward to see if there was a different in proportions for cars that did not stop for black pedestrians vs white pedestrians in Las Vegas. We found that there was, estimating the proportions cars not stopping to be 29 percentage points higher for black pedestrians than white pedestrians in Las Vegas. Our 95% confidence interval was a bit wide, being between 5.67 and 49.75 percentage points higher.

We also estimate, using the relative risk ratio, that black pedestrians are 3 times more likely to not have a car stop for them than white pedestrians in Las Vegas. The are 95% confident that our population value is between 21.21% and 8 times more likely. This interval, again, is rather large. Based on our analysis we seem to have significant evidence that whether a car stops may be dependent on race, specifcally in Las Vegas. However, the reason for this is unclear and more information analysis would be helpful to understand why.

2. Music and Food Choice

There was a study on the effect of background music volume on food purchasing habits in restaurants. The volume had to levels, low-volume(55-db) and high-volume(70-db). Food choices were rated at three levels: healthy, neutral, and unhealthy. Let's find out if the background music volume in restaurants has an effect on peoples food choices.

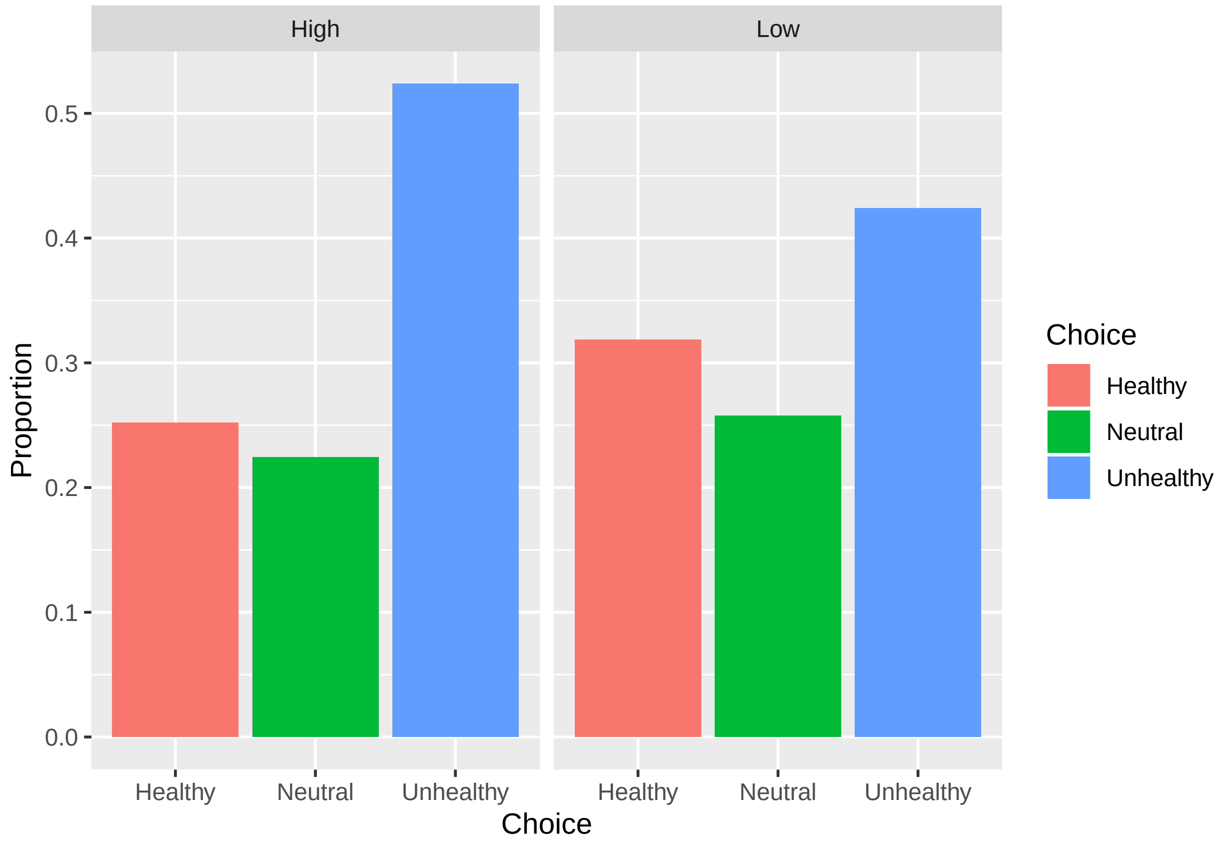
Here's a table of our observed data:

```
musicfood.counts
```

```
##
##         Healthy Neutral Unhealthy
##   High       64      57       133
##   Low        94      76       125
```

Data Visualization

```
ggplot(data = musicfood.cond.frame, aes(x=Choice, y = Proportion, fill = Choice)) + geom_bar(stat = "id
```



Based on the graph of the proportions of food choices separated by volume level, there seems to be a difference in unhealthy food choices between the volume levels. The difference in healthy and neutral choices based on volume are noticeable, but not as much as for unhealthy food choices.

Chi-Squared Test for Independence

Ho: Music volume and food choice are independent Ha: Music volume and food choice are not independent $\alpha = 0.05$

```
chisq.test(musicfood.counts)$expected
```

```
##
##         Healthy Neutral Unhealthy
##   High 73.10018 61.5337  119.3661
##   Low  84.89982 71.4663  138.6339
```

Our expected counts are at least 5 in each cell, so we will move on to perform the Chi-Squared Test for Independence

```r
chisq.test(musicfood.counts)
```

```
##
##  Pearson's Chi-squared test
##
## data:  musicfood.counts
## X-squared = 5.628, df = 2, p-value = 0.05996
```

```r
chisq.test(musicfood.counts)$residuals
```

```
##
##           Healthy    Neutral  Unhealthy
##   High -1.0643657 -0.5779577  1.2478975
##   Low   0.9876355  0.5362927 -1.1579366
```

After performing the Chi-Squared Test for Indeendence, we received a p-value of 0.06. We do not have significant evidence to reject the null, that music volume and food choices are independent.

Statistical Analysis

Based on a short but revealing analysis, it does not seem that there is a relationship between music volume and food choices. Our Chi-Squared Test for Independence did not yield significant evidence against the null hypothesis that volume and food choice are independent. Considering there is no evidence that there would be a signficant relationship between music volume and food choice in relationships, it does not seem purposeful to continue to analyze the data.