Final Exam: Take-Home Component

Taylor Blair

Math 141, Week 14

Instructions

- You have two consecutive hours to complete this exam. **Make sure to fill-in your start day/time and end day/time:**

Start day/time: 12/10/2020 9:32 AM (My birthday, also in a hospital, also have three not so fun drains coming out of me, and occasionally an IV)

End day/time: 12/10/2020 11:16 AM (I wasn't able to solve some of the last problems. So I am calling it a day and going to take a nap)

- Once you have completed the exam, turn in the knitted pdf on Gradescope.
- For this component, you are allowed to consult any materials from this course: the textbook, your notes, class handouts, lab assignments, worksheets, project assignments, RStudio cheat sheets, RStudio help files.
- You are not allowed to consult anybody (classmates, friends, family, teachers, random people on the street, etc.) for help on this exam. You must work on it alone.
- Do **NOT** post questions about the take-home exam to any Slack channels or anywhere else on the internet.
- Send technical questions to Kelly, Tom, and Jonathan as a joint Direct Message. For equity reasons, we will not answer conceptual or clarifying questions.
- Remember the Honor Principle.

The Prompt

Across the US recently there have been protests against systemic racism. Among other things, protesters have called for an end to discriminatory policing and have argued that black Americans are more likely to be stopped and searched by police. In this exam, we will use data on police traffic stops in San Francisco, California to examine a potential relationship between a driver's race and whether or not they are searched. We have data on traffic stops for June of 2016 and will assume that it is a representative sample for the city of San Francisco, generally.

Notes

1. A traffic stop is a temporary detention of a driver by police to investigate a possible crime or violation of law.
2. For all included graphs, make sure to include informative labels and titles.
3. For inference problems, feel free to use either simulation-based methods or probability model-based methods.

4. Pace yourself. Provide complete but also **concise** solutions. Providing details that are not asked for will make it difficult to complete the exam in 2 hours.

The Data

For each traffic stop in June of 2016 in San Francisco, California, we have:

- `search_conducted`: Whether or not the driver was searched
- `subject_race`: Race of the driver
- `date`: Date
- `day_of_week`: The day of the week
- `lat`: Latitude
- `lng`: Longitude
- `arrest_made`: Whether or not an arrest was made
- `citation_issued`: Whether or not a citation was issued
- `warning_issued`: Whether or not a warning was issued
- `outcome`: What the outcome was (arrest, citation, warning, or NA)
- `contraband_found`: Whether or not contraband was found on the driver (for the stops where `searched_conducted` = "yes")
- `reason_for_stop`: The reason for the traffic stop

```
library(tidyverse)
stops <- read_csv("/home/courses/math141f20/Data/sf_stops_june2016.csv")
```

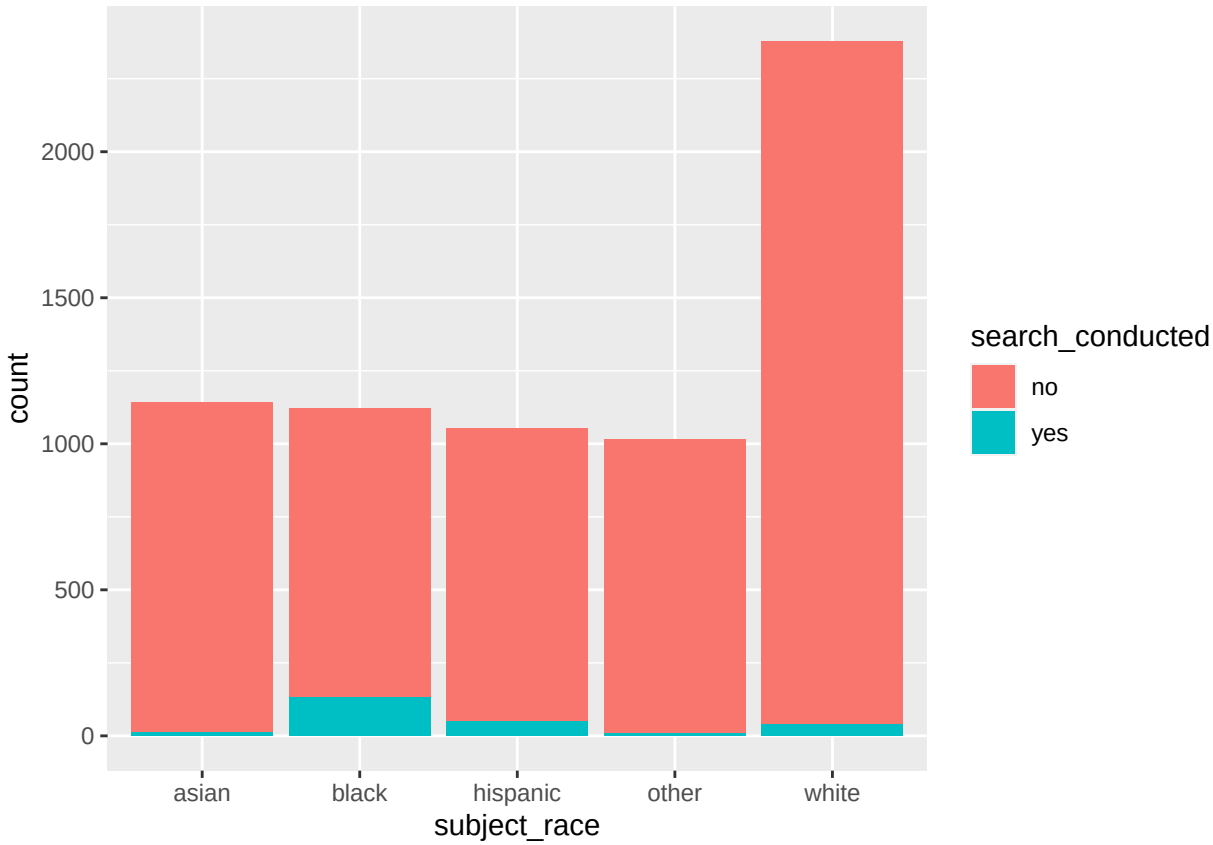
Problems

Problem 1

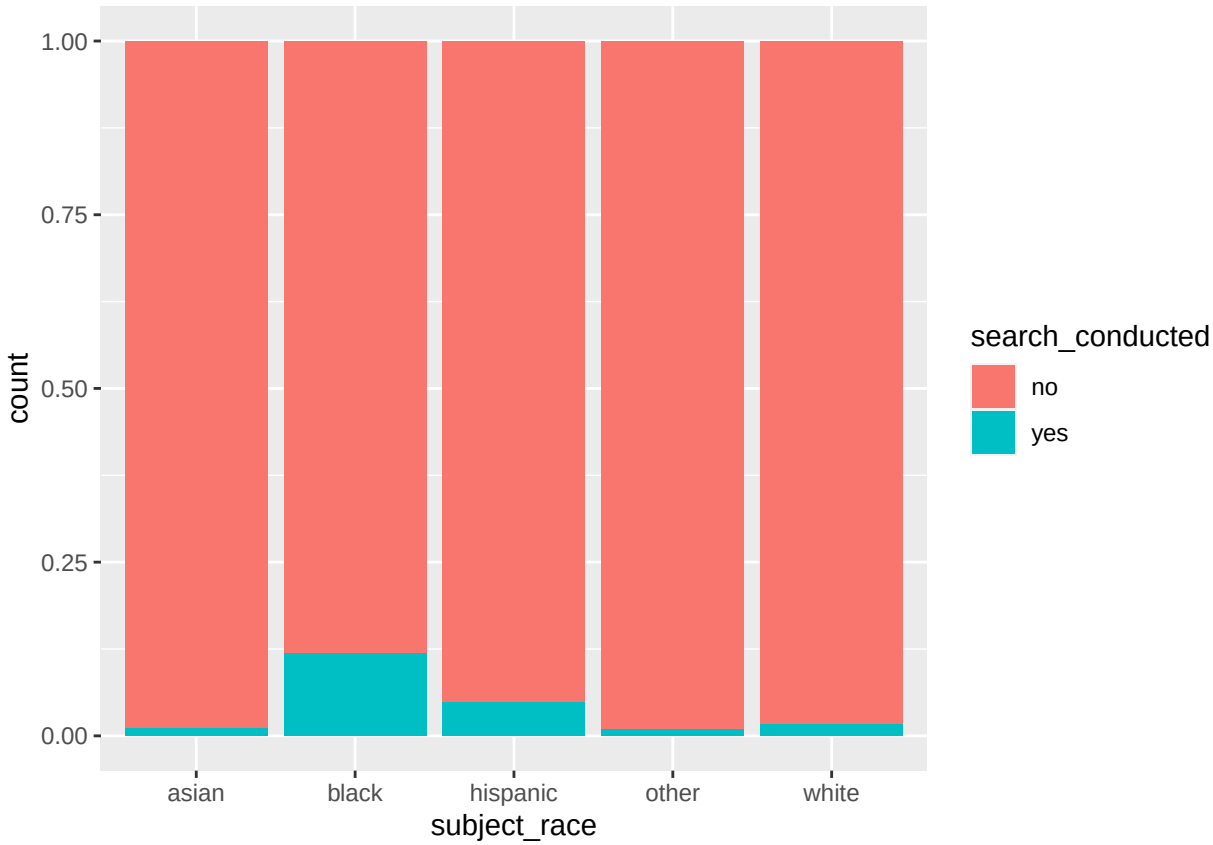
a. Create two bar graphs of `search_conducted` and `subject_race`:
   - One that includes counts.
   - One that includes conditional proportions.

Hint: For readability of labels, you may want to consider including a `theme(legend.position = "bottom")` layer.

```
ggplot(stops, aes(subject_race, fill=search_conducted)) +
  geom_bar()
```




```
ggplot(stops, aes(subject_race,  fill=search_conducted)) +
  geom_bar(position = "fill")
```



b. For each bar graph in (a), provide a useful conclusion that can't be easily drawn from the other bar graph.

- *The color scheme is searing my eyes*
- The bar graph is not proportional to the city of California, thus white individuals appear to have been stopped more
- The first graph shows that white individuals are stopped more py the police.
- That being said, the proportion of minority individuals that were searched is greater than white individuals in the second graph.

---

c. Create a table/data frame that contains the proportions searched or not searched by race. Sort the observations from the highest proportion to the lowest proportion. Which race has the highest rate of being searched? Which race has the lowest rate of being searched?

```
#pivot <- table(stops$subject_race, stops$search_conducted)

pivot <- stops %>%
  count(subject_race, search_conducted)

print(pivot)
```

```
## # A tibble: 10 x 3
##    subject_race search_conducted     n
##    <chr>        <chr>            <int>
##  1 asian        no                1130
##  2 asian        yes                 12
```

```
##  3 black        no                 987
##  4 black        yes                133
##  5 hispanic     no                1002
##  6 hispanic     yes                 51
##  7 other        no                1006
##  8 other        yes                 9
##  9 white        no                2340
## 10 white        yes                 39
```

```
cat("African American indivuduals are searched: ", 133/(133+987)*100, "%")
```

```
## African American indivuduals are searched:  11.875 %
```

```
cat("Hispanic indivuduals are searched: ", 51/(1002+51)*100, "%")
```

```
## Hispanic indivuduals are searched:  4.843305 %
```

```
cat("White indivuduals are searched: ", 39/(2340+39)*100, "%")
```

```
## White indivuduals are searched:  1.639344 %
```

```
cat("Asian individuals are searched: ", 12/(1130+12)*100, "%")
```

```
## Asian individuals are searched:  1.050788 %
```

```
cat("Other individuals are searched: ", 9/(9+1006)*100, "%")
```

```
## Other individuals are searched:  0.8866995 %
```

- Black individuals have the highest rate of being stopped at 11%
- The lowest goes to `other`

---

d. Conduct a hypothesis test to determine if race and whether or not a search is conducted are related. In your written solution, include
   - A test statistic and p-value
   - Whether or not test assumptions are met
   - Conclusions at $\alpha = 0.05$

```
chi <- chisq.test(table(stops$subject_race, stops$search_conducted))


cat("Test stastic of ", chi$statistic)
```

```
## Test stastic of  292.0426
```

```
cat("P-value of ", chi$p.value)
```

```
## P-value of  5.638069e-62
```

- Assumptions are met, there is more than thirty of each category for race
- This means that the data id proportionally different from
- Because our p-value is $5.6380689 \times 10^{-62}$, we have met the $\alpha = 0.05$. Thus we can reject the null.

---

e. State what a Type I error and a Type II error mean in the context of the hypothesis test conducted in (d).

- Type 1 *False Positive*
  - We declare that minority individuals are more likely to be searched by police when in reality the inverse is true
- Type 2 *False Negative*
  - Declared that there is no difference incorrectly

---

f. Give a potential confounding variable. Address how this variable differs between the explanatory variable groups and how it might have a potential effect on the response variable.

- We didn't look at the distibution of races in san francisco. We assumed that the total stops was proportional to the actual props of SF.

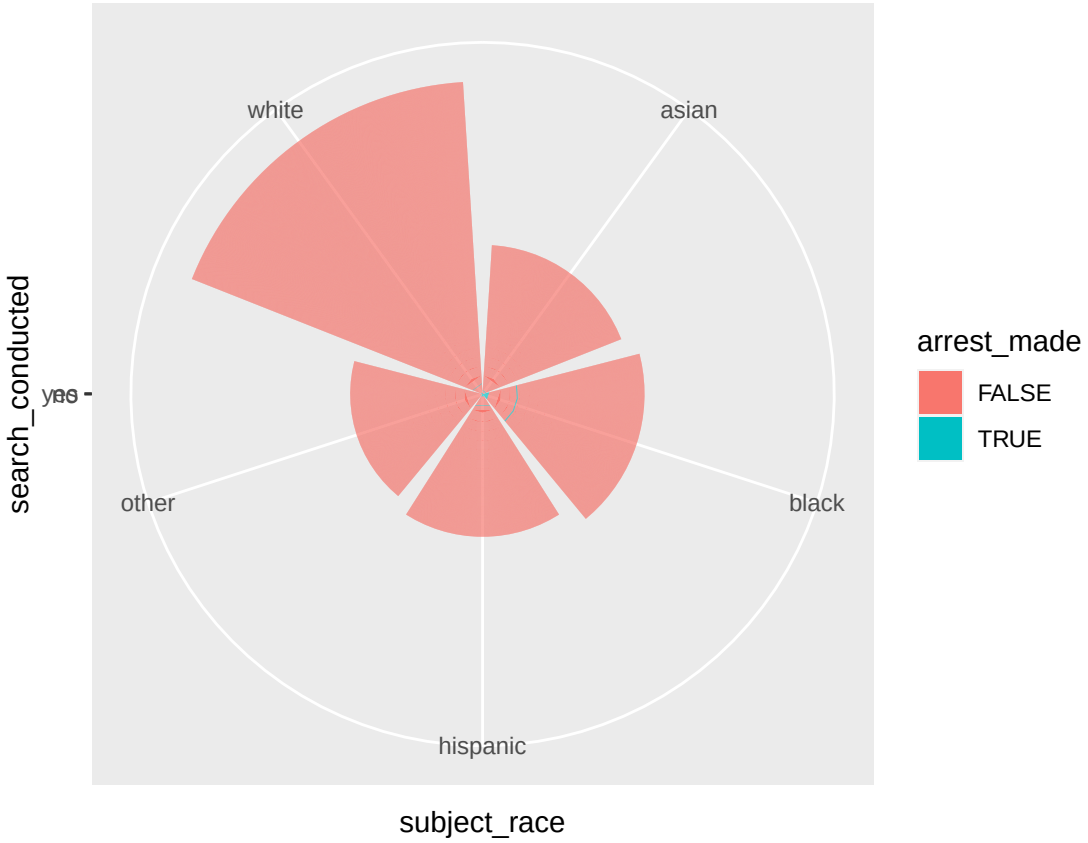
---

g. As we have discussed often, we live in multivariate world. Create a graph that explores the relationship between a driver's race and whether or not a search was conducted but also incorporates a third variable.

Using only your graph answer the following question: Does controlling for another variable seem to **increase**, **decrease**, or **not impact** the strength of the relationship between the driver's race and whether or not they are searched? Justify your answer.

```
# My goal is to make a round bar chart with half being search yes, and other no

ggplot(stops, aes(subject_race, search_conducted, fill= arrest_made)) +
  geom_col() +
  coord_polar()
```



- Tiny little ribbon of arrest made in cars when searches were made for black individuals

---

Problem 2

a. Create a data frame that only contains the stops where a search was conducted and where the driver was either black or white. (Use this dataset for the rest of Problem 2.)

```
white <- stops %>%
  filter(subject_race=="white")

black <- stops %>%
  filter(subject_race=="black")

prob_2 <- rbind(white, black)

sample_n(prob_2, 5)
```

```
## # A tibble: 5 x 12
##   search_conducted subject_race date       day_of_week   lat   lng arrest_made
##   <chr>            <chr>        <date>     <chr>       <dbl> <dbl> <lgl>
## 1 no               black        2016-06-09 Thu          37.7 -122. FALSE
## 2 no               white        2016-06-25 Sat          37.7 -122. FALSE
## 3 no               black        2016-06-08 Wed          37.8 -122. FALSE
## 4 no               white        2016-06-02 Thu          37.8 -122. FALSE
## 5 no               white        2016-06-17 Fri          37.8 -122. FALSE
```

```
## # ... with 5 more variables: citation_issued <chr>, warning_issued <chr>,
## #   outcome <chr>, contraband_found <chr>, reason_for_stop <chr>
```

b. Construct a 90% confidence interval and a 95% confidence interval for the difference in the proportions of contraband being found between white and black drivers.

```
quantile(black$contraband_found, probs=seq(0.025, 0.975), na.rm=TRUE, names=TRUE)
```

c. If there is no discrimination, researchers believe that searches should yield contraband at the same rate for each race. If searches of minority drivers yield contraband at lower rates, it may be suggestive of discrimination. Based on your confidence intervals, draw some conclusions about how the rates do or do not differ between blacks and whites?

- I am going to go out on a limb and say race is involved.

Problem 3

Suppose we want to estimate the proportion of traffic stops in Portland, OR where the driver is searched. We want to construct a 98% confidence interval with a margin of error bounded by 0.02. How many traffic stops should we sample?

```
cat("90%", ceiling((0.98^2*0.02*(1-0.02))/(0.04^2)))
```

```
## 90% 12
```

- Isn't the rule ten times the prop you want, so 980
