STAT 113: TESTING FOR ASSOCIATION BETWEEN CATEGORICAL VARIABLES

The phrase "driving while black/brown" (a reference to "driving while intoxicated") refers to the phenomenon of a black driver (or other driver of color) being stopped by police or, once a stop has occurred having their vehicle searched, without legitimate reason to suspect wrongdoing, but merely for the offense of "driving while black(/brown)". That is, it is concerns the practice of *racial profiling*. Racial profiling can occur at many stages during an interaction with law enforcement.

In this investigation, we will focus on the following question: **Once a stop has occurred, is there evidence that drivers of color are more likely to have their vehicle searched than white drivers?**

Armentrout, et al. (2007)[1] reports data on a variety of outcomes related to traffic stops by the Los Angeles Police Department (LAPD). Two of the variables recorded are (1) race the driver and (2) whether or not the vehicle was searched.

One question we could ask is whether the proportion of drivers who are searched (out of those who are stopped) differs across racial categories. Put another way, we are interested in whether the two variables (driver race and search status) are *associated* in the population.

We can state our null and alternative hypotheses generally as

$$H_0 : \text{there is no association between the two variables}$$
$$H_1 : \text{there is some association between the two variables}$$

If the variables were quantitative, we would assess association with correlation or regression, but what can we do with two categorical variables? We will develop a method to do this in this worksheet.

Date: November 30, 2017.

[1]Armentrout, M., Goodrich, A., Nguyen, J., Ortega, L., Smith, L., & Khadjavi, L.S. (2007). Cops and stops: Racial profiling and a preliminary statistics analysis of Los Angeles police department traffic stops and searches. Retrieved from http://www.public.asu.edu/ etcamach/AMSSI/reports/copsnstops.pdf

(1) The sample consisted of 6387 stops in total. Of those, the breakdown of driver race was as follows: 2336 Hispanic/Latinx (36.6%), 2190 white (34.2%), 1248 black (19.5%), 502 Asian (7.9%), and 111 of all other classifications combined ("other", 1.7%). In total, 882 (13.8%) of the stops involved a search by law enforcement.

Suggest a method to construct one randomization sample, assuming that that H_0 is true and there is no association between these two variables. Your method can use cards or some other random number generator. There is more than one possibility!

(2) The full dataset is summarized by the following contingency table.

	Hisp./Lat.	White	Black	Asian	Other	Total
Searched	510	109	240	16	7	882
Not Searched	1826	2081	1008	486	104	5505
Total	2336	2190	1248	502	111	6387

What statistics could you compute from this data that would give you a sense of whether the two variables are associated?

(3) Suggest a way to measure association using a single number for data like this. There are many possibilities, but it should give "more extreme" values when the association between race and search status is largest. What kinds of values count as "extreme" for your measure? Large values? Small values? Extreme values in either direction?

(4) One randomization scheme you might use is exactly analogous to the scheme we used to test correlations: use $n = 6387$ cards, each labeled as "searched" or "not searched". You would then shuffle them, and divide them into five piles, representing the race of the driver. How many cards of each type should you include?

(5) If you were to do this simulation, for each full simulation of a sample of $n = 6387$, you could compute the test statistic you came up with in question 3. This becomes one point on the randomization distribution. After repeating this whole process thousands more times, you would have a a collection of simulated test statistics constructed under the assumption that H_0 was true. How would you calculate a P-value from this distribution and your data?

(6) Many reasonable test statistics are possible. For example, we could compute the search rate for each driver race, and compute the variance of these rates. Large values mean that there are larger disparities across driver races. Or we could compute the absolute value of each pairwise difference of proportions and average these. Again, larger values mean larger disparities. Or, we can compute a χ^2 statistic, as we did to test goodness of fit for a single categorical variable. In this case since there are two variables, we observed and expected counts for each *combination* of values.

$$\chi^2 = \sum_{r=1}^{C} \sum_{c=1}^{R} \frac{\left(\mathsf{Observed\ Count}_{r,c} - \mathsf{Expected\ Count}_{r,c}\right)^2}{\mathsf{Expected\ Count}_{r,c}}$$

where r and c range over the rows and columns of the contingency table, respectively.

How can we find the expected counts in this case if H_0 is true? (Hint: use the fact that "no association" means the percentage of searches within each race is the same)

(7) To find expected counts under H_0, we assume that the search rate is 13.8% for *each* driver group. So, of the 2336 total Hispanic/Latinx drivers stopped, we expect that $0.138 \times 2336 = 322.37$ will be searched, the other $0.862 \times 2336 = 2013.36$ will not, and so on for the other groups. Note that these are not whole numbers; but that is okay, since they represent averages across many samples. All together, we have the expected counts shown below:

Expected	Hisp./Lat.	White	Black	Asian	Other	Total
Searched	322.37	302.22	172.22	69.28	15.32	882
Not Searched	2013.63	1887.78	1075.78	432.72	95.68	5505
Total	2336	2190	1248	502	111	6387

Compare to the observed counts (repeated below for convenience):

Observed	Hisp./Lat.	White	Black	Asian	Other	Total
Searched	510	109	240	16	7	882
Not Searched	1826	2081	1008	486	104	5505
Total	2336	2190	1248	502	111	6387

In StatKey, go to "χ^2 Test For Association", and click "Edit Data". Visit the following URL to get the data table which you can cut and paste into StatKey:

`http://colindawson.net/data/racial_profiling.txt`

What is the sample χ^2 value? Construct the randomization distribution and compute a P-value. What is your conclusion?

(8) Now that we have concluded that there is evidence of an association, we would like to be able to say something about which categories are contributing to the association. Click "Show Details" to see which cells of the table have particularly large standardized deviations from our expectations (i.e., look at the contribution to the χ^2 statistic from each cell). Which cells have the biggest discrepancies?

(9) We can also use a theoretical approximation in place of the randomization distribution. As with the Goodness of Fit test, the χ^2 statistic has a χ^2 distribution with some number of degrees of freedom, when the null is true. In the one-variable case, the degrees of freedom was one less than the number of categories. Here, we have a 5×2 table of counts *Having fixed the row and column totals*, how many cells do we need to fill in using the data before the rest are determined for us? This number is the degrees of freedom.

(10) Find a P-value using the appropriate χ^2 theoretical distribution in StatKey. What is your conclusion in context?