David Kinney

dsc530 Spring 2019

**Date**: June 2, 2019

**Instructor**: Professor Catie Williams

# Statistical/Hypothetical Question

Given the NHTSA statistics for national vehicle fatalities in 2017, is it possible to identify the predominant causes? Conversely, is it possible to identify variables that do not reflect a direct cause?

# Outcome of your EDA

Generally speaking, there was very little evidence that would lead to predominant causes of vehicular fatalities. For example, most accidents took place in broad daylight, and clear weather conditions. It is just an opinion, but perhaps this might be due to the fact that most drivers are on the roads during the day when the weather is not inclement. Therefore, I feel it makes sense for me to conclude that the first part of my hypothesis is not provable, however the second part is more in line with what I’ve discovered.

# What do you feel was missed during the analysis?

The FARS data is huge, encompassing a number of individual tables (the FARS guide is available from my GitHub repo). I focused on just two: **accidents** and **person**. I have little doubt there is a plethora of useful data in the remaining tables that might lead to entirely different conclusions when not locked in to just two of the datasets.

# Were there any variables you felt could have helped in the analysis?

Pretty far into the process I came across data on alcohol and drug statistics. This could very well be an analysis project of its own. I might investigate this approach over the summer.

# Were there any assumptions made you felt were incorrect?

Sometimes slicing and dicing the data and the comparing and contrasting the resulting visualizations borders on a magic trick. One particular such example came about when visualizing the hour the accident took place to the day of the week. Plotting just those two variables outright paints one picture. However, when I compared probability mass functions for accidents on Wednesdays and accidents on Saturdays, an entirely different picture

emerged: Wednesday PMF is highest around rush hours, while Saturdays spike between midnight and 2am. I would never have reached that conclusion without slicing up the data differently.

# What challenges did you face, what did you not fully understand?

The biggest challenge was that I realized too far into the project that these FARS datasets did not lend themselves to this type of analysis. The vast majority of the variables and not continuous. It therefore became a bit of a struggle to create scatter plots and so forth.

I still feel overwhelmed with all of the different statistics available. I find it intuitive to graph variables and draw conclusions from this visual feedback. But keeping straight the large number of coefficients, correlations, P value, R-squared, and the two dozen others I haven’t mentioned is still a huge challenge. I plan to spend a good deal of the summer going back and taking each of them, one by one, and wrapping my head around them (I’ve become a big fan of Microsoft OneNote ® and will likely build a nice set of notebooks).

Lastly, Python for statistical analysis is a struggle for me. I kept typing in commands and wondering why they were throwing errors, only to realize I had typed in an R command. Another personal shortcoming I will work on over the summer break.