Capstone Project 1: Inferential Statistics

The two primary methods used in this inferential statistics section are Bayesian inference and bootstrapping.

As a recap: our dataset holds information on opioid prescriptions and opioid overdose deaths across every county in the United States between 2006 and 2012. During the data wrangling section of the project, I combined this data with demographic data from the 2010 Census for each county. The goal of the project is to discover the relationship between prescription rate and death rate, as well as to predict an individual's likely rates given their background (location, age, sex, etc.). For reference, the data on opioid overdoses was acquired from https://wonder.cdc.gov/ and the demographic data was from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml .

Because our dataset was missing death data from many counties, I needed to find some way to predict these values. I determined that imputing them with some aggregate statistic (mean, mode, median) was not the appropriate approach, considering significant differences between counties in terms of population size and characteristics. I decided to create a machine learning model that would be able to use a county's demographic data to predict its likely death total. For this model I used a Lasso regression in order to assure regularization and positive values in our result.

According to our data visualizations, there should be a fairly linear relationship between death rate and prescription rate. To evaluate this relationship in closer detail, I created a metric called "pain ratio", which is simply the ratio of death rate to prescription rate. The purpose of the metric is to define how deadly opioid prescriptions are - the higher the pain ratio, the more likely someone is to die from having a prescription.

Creating a histogram of pain ratios for the entire dataset shows that the mean ratio is about 0.1, but the median is about 0.05 and the mode is even smaller. This demonstrates that for the majority of counties, opioid prescriptions are not very deadly. The distribution of values appeared to roughly match a gamma distribution. I simulated a gamma distribution of random continuous variables using Scipy, in order to compare to the observed data. This simulated model has a similar shape to the observed distribution, although it appears to be less positively skewed.

Next, I used a Markov chain Monte Carlo method from PyMC3 to simulate the data with Bayesian inference. Because this is a gamma distribution, the parameters alpha and beta can also be simulated with an exponential distribution. I took 10000 samples of this model, then discarded the first half (the “burn-in period”) because an MCMC run starts from a random position and therefore takes some time to find the actual peaks of the distribution. Plotting this new simulation shows a model with a much wider distribution than the observed data - in fact, less than 5% of the means of simulated pain ratios were as small as the observed mean. If the null hypothesis is that opioid prescriptions are fairly deadly, then the small p-value from the Bayesian model allows us to reject the null hypothesis: opioid prescriptions are not likely to be more deadly than observed.

The next section of inferential analysis evaluates the differences between men and women. For this, I decided to implement bootstrapping (randomly sampling with replacement). Our null hypothesis is that there is no difference between men and women, while the alternate hypothesis is that a difference exists. Assuming the null hypothesis, we have to “shift” the values of the groups to simulate no difference between them. This is done by subtracting the mean of the group from each point in the group, then adding the mean of the combined groups.

Using bootstrapping, I tested this hypothesis for three variables - death rate, prescription rate, and pain ratio. These simulations yielded p-values, respectively, of 0.45, 0.66, and 0.82. Compared to the typical alpha value of 0.05, all of these results fail to reject the null hypothesis. We can therefore conclude that there are not significant differences in opioid prescriptions or opioid overdoses between men and women.

https://github.com/mdubow/Springboard-Data-Science-Career-Track/blob/master/Capstone-Project-1/Inferential-Statistics/Capstone%201%20Inferential%20Analysis.ipynb