Capstone 1: Predicting Prescriptions, Inferential Statistics

Rika Yatchak

January 25, 2019

1 Introduction

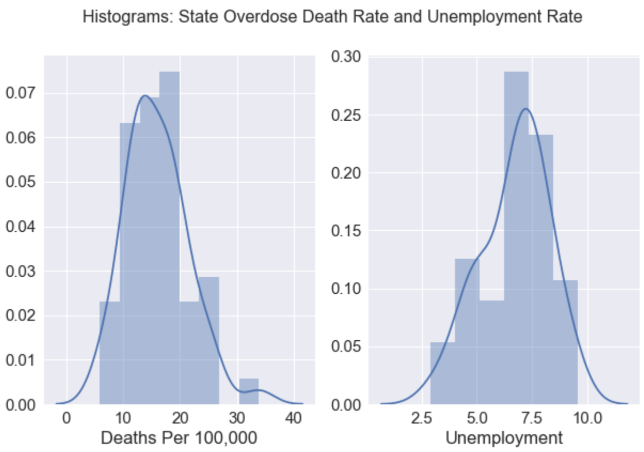
In this document we perform some inferential statistics on the Kaggle Opioid Prescriber Dataset for my Springboard Capstone project. To view the code used to generate the results in this report, see the corresponding Jupyter Notebook.

2 Correlations

We begin by evaluating correlation between the opioid overdose death rate per 100,000 in each state and two other variables: state population and state unemployment rate.

2.1 Overdose Death Rate vs. State Unemployment Rate

First we take a look at the histograms for Deaths Per 100,000 and Unemployment rate. Neither distribution appears to be normal, so we'll use the Spearman-r test for correlation. We will take $\alpha = .1$ as our critical value. We calculate the Spearman-R correlation coefficient r and test the null hypothesis that $r = 0$, that is, that the values are not correlated.

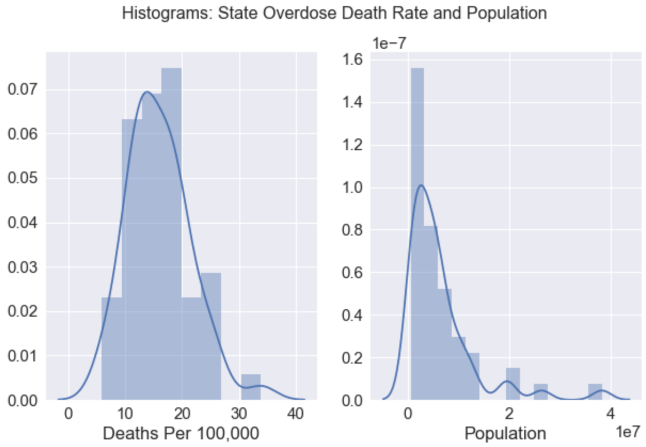


Carrying out the test yields $r = 0.265$ and $p = 0.063$. We reject the null hypothesis and conclude that the values are correlated. However, the correlation is not very strong, as was already evident in the scatterplot visualization performed while creating the data story.

The p-value here corresponds to a two-sided t-test, where the t-statistic is $t = r\sqrt{\frac{n-2}{1-r^2}}$, where r is the Spearman correlation coefficient, and n is the sample size. The statistic t is approximately distributed as a Student-t distribution with $n - 2$ degrees of freedom.

2.2 Overdose Death Rate vs. State Population

As before, we generate a histogram to look at the distribution of the data.



Population is decidedly non-normal, but this isn't much of a surprise. Using the Spearman r-coefficient test with $\alpha = .1$ as before, we have $r = -0.044$ and $p = 0.762$. We fail to reject the null hypothesis and conclude that there is no statistically significant correlation between state population and state opioid overdose death rate.

3 Distinguishing Between Providers Based on Drug Prescriptions

Our TSNE visualizations indicated that specialist providers of certain types can be distinguished from one another based on their *prescription series*: the series is indexed by drugs, and the value is the number of times the provider prescribed this drug in 2013.

The fact that we can distinguish prescribers in this way indicates that there could be some drugs which tend to be prescribed much more often by certain specialists than by others.

We'll explore this by taking a look at psychiatrists and orthopedic surgeons. We aggregate their

4 Psychiatrists versus Orthopedic Surgeons

In total, there are 691 providers with a Psychiatry specialty and 575 providers with an Orthopedic Surgery specialty in the dataset. We look at the drugs prescribed in both specialties, and then the mean amount prescribed by each provider (only the head of the dataframe is reproduced here):

Specialty	Orthopedic Surgery	Psychiatry
Drug		
TRIAMTERENE.HYDROCHLOROTHIAZID	0.433043	0.094067
SYNTHROID	0.173913	0.073806
DULOXETINE.HCL	0.387826	40.531114
HYDROCORTISONE	0.052174	0.028944
TIMOLOL.MALEATE	0.097391	0.000000

We notice, for example, that Duloxetine HCL is much more likely to be prescribed by a psychiatrist than an orthopedic surgeon.

We want to carry out a two-sample test to compare the means amount of a certain drug prescribed by the two different populations (psychatrists and orthopedic surgeons). We notice that the sample standard deviation is 5.426 for orthopedic surgeons, and 83.05 for psychiatrists. We don't know the true standard deviation of the populations, but it's certainly not fair to assume that they're equal. That means we shouldn't use a two-sample t-test here as the variance equality assumption is violated. We can use Welch-Aspin t-test instead.

Null Hypothesis: the mean amount of duloxetine hcl prescribed by psychiatrists and orthopedic surgeons is the same. We will reject the null hypothesis with the critical value $\alpha = .1$. We obtain a t-statistic $T = 12.67$, which yields $p = 2.837E - 33$. We reject the null hypothesis.

We have no domain knowledge, but we can already guess that Duloxetine is probably a specialist psychiatric drug. It turns out that Duloxetine is a widely-prescribed drug mostly used to treat depression and anxiety, as well as chronic pain disorders. It's no surprise that psychiatrists prescribe this drug much more often than orthopedic surgeons.

Testing mean hydrocortisone prescription in exactly the same way, we obtained a t-statistic $T = -.485$ and $p = 0.628$. We fail to reject the null hypothesis and conclude that both orthopedic surgeons and psychiatrists tend to prescribe smaller (but nonzero) amounts of hydrocortisone, at about the same rate. Given that hydrocortisone is used to treat a variety of ailments, this conclusion seems reasonable.

5 Distinguishing Between Generalists

Although the TSNE visualizations suggest that family practice and internal medicine providers are probably not as easily distinguished for one another, we do see some drug prescribing differences.

In total, there are 3 194 internal medicine practitioners in the dataset, and 2 975 family practice providers. Using the same procedure as before, we took a look at the observed mean prescriptions for the various drugs that were prescribed by both types providers. The mean prescriptions for many drugs were quite similar, but some drugs had differences that seemed like they could be statistically significant.

As before, we carried out a Welch-Aspin t-test with $\alpha = .1$. Null hypothesis: the mean amount of synthroid prescriptions written by family practice providers is the same as the mean amount of synthroid prescriptions written by internal medicine providers. This yields a t-value of $T = -3.859$ and $p = 0.0001$. We reject the null hypothesis.

6 Conclusion

As our TSNE visualizations suggested, we can distinguish between specialist prescribers based on the amount of certain types of drugs that they prescribe. This suggests that a logistic regression task to predict prescriber specialty from amount of drugs prescribed in a year should have a reasonable amount of success, at least for specialists.

Moreover, we saw that state opioid overdose death rate was weakly correlated to the state unemployment rate. The lack of groundbreaking results in the opioid overdose death rate statistical analysis is unsurprising: since the opioid epidemic is most commonly studied on a county-wide basis. However, access to cause of death information on the county level is (justifiably) difficult to obtain and generally only available to registered public health researchers. Therefore, this project will concentrate on predicting prescriber specialty based on drug prescriptions.