# Project Stage - III (Distributions and Hypothesis Testing)

#### Task 1: Distribution Analysis (50 pts)

* Member: (50 pts)
  + **M1.1** Compare NC and KY on Opioid Mortality - 2019 Data (20 pts)
* Create histograms for NC and KY for Opioid Mortality (Normalized Mortality Rate)
* Merge them into a single graph
* Plot mean lines for both the histograms

**Solution:** In order to do this, I have used super dataset. I filtered the data for NC and KY state and saved it as separate data frames. Then plot the histogram for each of them.

Merge 2 histograms on one plot and added the mean lines. Below is the screenshot for the same.

Chart, histogram

Description automatically generated

**M1.2** Evaluate a distribution for the Normalized Mortality Rate (20 pts)

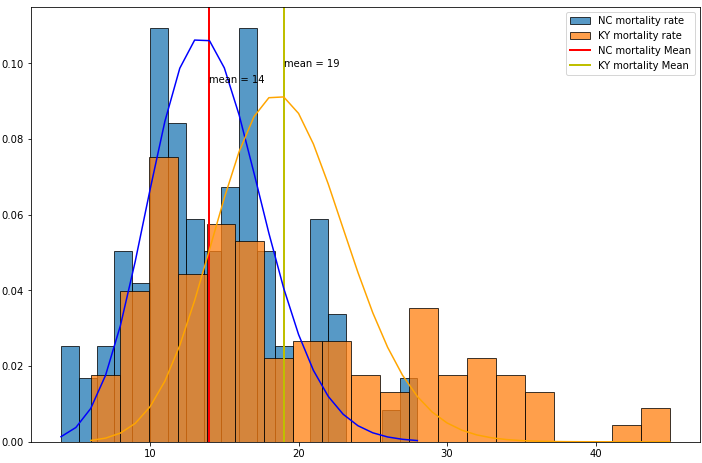
* Choose a distribution for Normalized Mortality Rate
  + Provide explanation of your choice
* Develop distribution estimator with - Method of Moments (MoM), Maximum Likelihood (MLE), and Kernel Density Estimation (KDE)
* Plot the distribution estimations over the histograms
  + Each should be color coded on the graph and described for their parameters
  + Discuss which estimator works the best and why

**Solution:** Now, we will do distribution estimators for NC and KY for mortality rate.

The mortality rate values are discrete. So, we have to figure out some discrete distribution. There are many discrete distributions like monomial, binomial, Poisson, and geometric distribution. But we are going to use Poisson distribution. The reason for selecting Poisson is that this distribution is time dependent and gives probability of observing number of successes in given interval of time. We have Year information with us and want to figure out no. of deaths in given year which can be done by Poisson distribution.

I have also tried other distributions like binomial, but it wasn’t giving the best results and wasn’t covering the histogram fully. So, I settled with the Poisson distribution. Below are the screenshots for MOM, MLE and KDE plots for NC and KY state.

**MOM:**



**MLE:**

Chart, histogram

Description automatically generated

**KDE:**

Chart, histogram

Description automatically generated

**M1.3** Select the top two states identified in Stage I and recreate the M1.2 task (10 pts)

**Solution:** The top 2 states are TN and TX identified in stage I. Plotting estimators for them.

**MOM:**

Chart, histogram

Description automatically generated

**MLE:**

Chart, histogram

Description automatically generated

**KDE:**

Chart, histogram

Description automatically generated

#### Task 2: Hypothesis Testing and Regression (50 pts)

* Member: (50 pts)
  + **M2.1** Formulate Hypothesis for 5 identified variables in Stage 1 and test the hypothesis (25 pts)
    - For example, if your variable was Premature death raw value, divide the observations into two categories, high and low
      * Compare the distributions for a hypothesis test.
    - Formally state the Null and Alternative Hypothesis
    - Define the type of hypothesis and the thresholds
    - Conduct the test and discuss the results for all 5 variables

**Solution:** I used paired t-test to perform hypothesis testing. It checks for similarity/differences in same variable at different intervals in time.

My 5 variables are:

* opioid dispensing rate
* unemployment rate
* drug overdose
* insufficient sleep
* excessive drinking

### **Hypothesis Test 1.** Norm Deaths vs Opioid dispensing rate

**Null Hypothesis H0:** There is no significant difference between Norm Deaths value generated by high and low opioid dispensing rate.  
**Alternative Hypothesis H1:** Norm Deaths values generated by high and low opioid dispensing rates differ significantly.

I divide the data into 2 ranges, low and high based on median (quantile = 0.5) and check for Norm\_Deaths value in both the datasets if its same/different using ttest\_rel() which is to do paired t-test in python. I have showed here only for one variable. In the notebook I have done it for all 5 variables. I reject/fail to reject null hypothesis based on p-value. If its less than significance level we can reject it.

* **M2.2** Perform linear regression to discover patterns (25 pts)
  + Perform a linear regression between Normalized Mortality and Opiod\_Dispensing\_Rate
    - Normalize the Opiod\_Dispensing\_Rate to the population
  + Perform multiple linear regression model with your 5 variables and Opiod\_Dispensing\_Rate
    - Evaluate the results and describe them.
  + Test non-linear model with the 5 + 1 variables (n=2,3,4)
    - Evaluate the results and describe them.

**Solution:** Linear regression is easy to calculate between mortality rate and dispensing rate.

The slope and intercept value is used to determine the relationship between them. Positive slope means positive relationship else negative relationship. The r-square value determines the fittedness of model to the distribution.

For multiple regression, I have used 5 variables described above and the output represents the slope values and intercept. The summary() is used to get summary of model. p-value and r-square value is important factor to check in summary.

For non-linear regression, I have used both sklearn and stats library and check for polynomial degree = 2, 3, 4. The same statistics are checked (r-square, p-value and slope). Based on r-square value determines if linear model or non-linear model is good. In non-linear also, which is a good fit (degree = 2, 3 or 4) can be checked by r-square value.