**Project Stage - III (Distributions and Hypothesis Testing)**

**Raveena Arasikere Rakesh**

1. In task-1 we first read the dataset,
   1. Calculated the mean and median for both the NC and KY data.
   2. Plotted both the NC and KY data in a single plot indicating the mean of both the states. Chart, histogram

      Description automatically generated
2. In task 2,
   1. We analyze the type of distribution across the normalized mortality rate. Since, the values are rounded, we categorize them as a discrete distribution. We assume that the data follows a **Poisson distribution.**
   2. In a Poisson distribution, the calculated mean and standard mean are going to be the same across the distribution. Hence both the MOM and MLE are going to be the same, and this was verified by plotting a graph. When we plot the graph both the MOM and MLE curves overlap. Hence, we regard both as the same.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

* 1. We plot the KDE using the plot.kde() function.

Chart, histogram

Description automatically generated

* 1. Next, we plot all the 3 curves in a single plot to analyze which curve fits the distribution the best.

Chart, histogram

Description automatically generated

##### From the above generated curve, we can see that method of moments and the most likelihood estimate remains the same for both a Poisson distribution and hence it is proved using the graph, also both act as a good fit for the graph and generates a best fit curve and hence we choose mom/mle as the best fit curve for the NC data

* 1. Next, we perform the same operations on the KY data to analyze which curve fits the best for a Poisson distribution.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

* 1. Plotting both MOM/MLE and KDE in a single plot.

Chart, histogram

Description automatically generated

##### **From the above generated curve, we can see that KDE generates a best fit curve and hence we choose KDE as the best fit curve for the KY data.**

* 1. **Next, we plot both the NC and KY data in a single graph along with their mean values. (Here MOM/MLE is the best fit curve for the NC data and KDE is the best-fit curve for KY data-which is indicated by the legend)**

Chart, histogram

Description automatically generated

* 1. Next, we perform the same analysis on the top-2 states with highest mortality rate identified in stage-1. Those states were WV and NM.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

##### *From the above graph we can see that MOM/MLE can be considered as the right fit curve for wv data*

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

* **From the above curve we can see that MOM/MLE performs better and can be considered a bestfit**
  1. Next, we plot both WV and NM data into a single graph, along with their mean values.

Chart, bar chart, histogram

Description automatically generated

1. In task-3 we formulated a hypothesis on the selected 5 variables in comparison with a target variable- Opioid mortality rate. Since it is a single variable comparison with a target variable, we performed a two sample t-test to either accept or reject the stated hypothesis.

##### *The 5 variables selected in the stage-1 of the project were-*

1-Excessive drinking raw value  
2-Sexually transmitted infections raw value  
3-Preventable hospital stays raw value  
4-Mental health providers raw value  
5-Premature death raw value

* First, we need to normalize each of the column with population per 100,000 so that the scale of comparison is the same.
* Next, we need formulate the hypothesis and then perform a two-sample t-test because we will be analyzing two independent variables.
* Calculate a threshold for each of the selected 5 variables.
* We set a standard way to set the threshold which is calculated as (min+max)/2.
* Values that are less the threshold are set to LOW and the values greater than the threshold are set to HIGH.
* Once a threshold is set, we then perform a two-tail ttest to see check if we can accept or reject our NULL hypothesis.
* We compare the HIGH values across Normalized Mortality rate with the HIGH values across the selected 5 variables.
* Next, We compare the Low values across Normalized Mortality rate with the Low values across the selected 5 variables.

Final analysis and result:

**1) For excessive drinking raw value:**

*- H0: Across any state in the US, mean of excessive drinking rate similar to mean of the opiod mortality rate*

*- H1: Across any state in the US, there is significant difference between the means of excessive drinking rate and opiod mortality rate*

- The threshold set is (max+min)/2=0.13396205075342518

- The hypothesis test performed is two-sample T-test

- Result: Reject the Null hypothesis.

**2) For Sexually transmitted infections raw value:**

*- H0: Across any state in the US, mean of Sexually transmitted infections raw value is similar to mean of the opiod mortality rate*

*- H1: Across any state in the US, there is significant difference between the means of Sexually transmitted infections raw value rate and opiod mortality rate*

- The threshold set is (max+min)/2=665.0270501638447

- The hypothesis test performed is two-sample T-test

- Result: Reject the Null hypothesis.

**3) For Preventable hospital stays raw value:**

*- H0: Across any state in the US, mean of Preventable hospital stays raw value is similar to mean of the opioid mortality rate*

*- H1: Across any state in the US, there is significant difference between the means of Preventable hospital stays raw value rate and opioid mortality rate*

- The threshold set is (max+min)/2=6557.256221900106

- The hypothesis test performed is two-sample T-test

- Result: Reject the Null hypothesis.

**4) For Mental health providers raw value:**

*- H0: Across any state in the US, mean of Mental health providers raw value is similar to mean of the opiod mortality rate*

*- H1: Across any state in the US, there is significant difference between the means of Mental health providers raw value rate and opiod mortality rate*

- The threshold set is (max+min)/2=0.008067995525588043

- The hypothesis test performed is two-sample T-test

- Result: Reject the Null hypothesis.

**5) For Premature death raw value:**

*- H0: Across any state in the US, mean of Premature death raw value is like mean of the opioid mortality rate*

*- H1: Across any state in the US, there is significant difference between the means of Premature death raw value rate and opioid mortality rate*

- The threshold set is (max+min)/2=11141.092263818999

- The hypothesis test performed is two-sample T-test

- Result: Reject the Null hypothesis.

4. We performed linear regression, multi-linear regression, and polynomial-regression on the selected 5 variables-in comparison with the Opioid mortality rate.

Chart, scatter chart

Description automatically generated

* **From, the above generated graph we can see that there a linear relationship between opioid morality rate and the opioid dispensing rate.**

##### j. *To perform a multiple linear regression model, we fit the data with the 5 variables and use the opioid dispensing rate as the target variable.*

Chart, line chart

Description automatically generated

RMSE score for multiple linear regression: 14.760761738406782

R-squared value: 0.24644317140642225

* The square root of the variance of the residuals is represented by the RMSE error- which is approximately 15, we can say that the model performed well but not did a great job in predicting the opioid mortality rate when the 5 variables are given.
* Since, R2 value is less than half i.e., less than 0.5 we can say that the model does not do a good job in explaining the proportion of the variance between the dependent and independent variable.

k. Performing Non-linear polynomial regression on the 5 variables in comparison with the opioid mortality rate.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

* For degree 1:
  + RMSE score for multiple linear regression for degree:1 14.760761738406782
  + R-squared value for degree:1 0.24644317140642213
* For degree 2:
  + RMSE score for multiple linear regression for degree:2 12.485227815405297
  + R-squared value for degree:2 0.4608725907118788
* For degree 3:
  + RMSE score for multiple linear regression for degree:3 7.85071603644558
  + R-squared value for degree:3 0.7868344556440677
* For degree 4:
  + RMSE score for multiple linear regression for degree:4 6.765188282082652
  + R-squared value for degree:4 0.8417082627869036

###### ***From the above generated results, we can see that the RMSE error is the lowest for degree-4 and the r-squared value is also the highest for degree-4 polynomial regression. Hence, we can say that the model performs the best for the degree-4 polynomial regression.***