# Stage 3 Report

**Task 1: Distribution Analysis**

In this task the two states death values were merged in order to create a concatenated histogram.

Chart, histogram

Description automatically generated

Inference:

* From the graph we observe that the distributions overlap each other approximately symmetrically.
* We can say that NC and KY distributions are same i.e the death rate in both the states is approx. the same.
* Since, death is a discrete variable the distribution should be discrete. I am considering Poisson distribution for this data as death is related to time and we are trying to find the probability of deaths in a given interval of time say 1 month, 2 month , 3 month and so on. Since the dataset has the year wise data we will be checking yearly what was the deaths rates in each of the states.
* The data is looks like normally distributed with 2 peaks, one at zero and the other the mean of the deaths for each state.The zero tells us that the death values went on declining in the later stages of the year.
* With mean of NC and KY equal to 624.65 & 900.30 tell us that the deaths from 1999 to 2020 is 624 and 900 in both the states.

**M1.2 Evaluate a distribution for the Normalized Mortality Rate**

**Choose a distribution:**

The Poisson distribution is a discrete distribution that measures the probability of a given number of events happening in a specified time period. Here we are checking the probability of deaths per year. I tried to use other distributions for discrete like binomial, but the distribution didn’t fit well as in binomial we are checking the probability of getting a variable as success. Thus, poisson fits the best for the dataset that we have.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Inference:

Essentially an estimate of the mean and variance of the distribution calculated using the sample data is what the estimator corresponds to and this can be generated using different paramaters. Based on the type of distribution parameters are chosen.

1. MOM

* Since we are checking the death rate with time I am using poisson distribution to find the estimator. In MOM, the estimator parameter for MOM is lamda which is nothing but the mean of the deaths column in our dataset.
* For NC and KY, we see that the probability of getting the mean is 0.10 and 0.9.
* The NC histogram fits well with the data but KY seems not to best fit the distribution.
* Thus, we check for other methods where the data can fit well.

2. MLE

* In MLE we try to maximize the theta parameter. Since the distribution chosen in MLE is lamda which is again the mean of the deaths column. We try to locate the mean and see the probability of getting that mean in the distribution. This way we are trying to maximize the MLE parameter that is the mean.
* IN NC and KY, The NC distribution fits well with the MLE but for KY the its found to be move towards right and doesnt fit the distribution well.
* Here the MLE should be in sync with the each both the dataset to call it the best estimator.

3. KDE

* In KDE there is no parameter to estimate we directly use the KDE function to see how the plot looks for the data.
* We see that the plot fits well and covers most of the data like a normal distribution and this could be the best estimator for the datasets.

Inference for TN and TX states

Looking at the above graphs, we trying to understand the distribution which estimator fits the data best. If we observe all the three methods i.e MOM, MLE and KDE all of them fit the data well.

MOM:

* The probability of finding the mean of Texas is 0.14 and that of Kentucky is 0.9.
* Since the data fits well with the poison distribution and the variable is a discrete variable, we will choose lambda as estimator parameter. This is nothing but the mean of the deaths for both the states.

MLE:

* In MLE, the probability of mean for TN and TX states are 0.14 and 0.9. Since we are trying find the best parameter for theta we use lambda which is nothing but the mean of deaths.
* The distribution using MLE estimator fits best and poisson distribution works very well for this data.

KDE:

* This is a non-parametric estimator and also fits the data well. In case of parametric distribution which can obtained using seaborn plot function, the bandwidth has to be adjusted in order to obtain a smoother curve.

All three estimators fit the data well compared to the NC and KY states.

**Hypothesis Testing and Linear Regression**

The variables I am considering for hypothesis are as follows:

\*Drug overdose deaths raw value

\*Injury deaths raw value

\*Adult smoking raw value

\*Firearm fatalities raw value

\*Insufficient sleep raw value

Formally state the Null and Alternative Hypothesis and Define the type of hypothesis and the thresholds. Hypothesis is supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation. Also, its necessary to test this to carry out the study of a sample to variables. Thus NUll hyporthesis come into picture.

Null Hypothsis (H0) : It states that there is no significant difference between the specified population. Alternative Hypothesis (H1) : The alternative is opposite of null hypothesis stating that there is relation ship between the variables.

The types of hypothesis are -

One tailed Test - performed to find the significant difference between the variables.

Two tailed test - performed to check if the sample mean is less than or greater than the population mean.

There are different threshold that will help us understand if a null hypothesis should be accepted or rejected like pvalue

if the pvalue is less than 0.05 then we reject the null hypothesis.

There are other functions like chi-contigency to check relation between two variables that are independant. P-Paired Test to check the if the data of interest differs with two different categories like checking weight of a person before the disease attacked him and then checking the weight after recovering. The Null hypothesis here is the weight remains same before and after.

With the example above we will try to check for other 5 variables.

I grouped the data by state and created a data frame that gives the mean of each state for each variable. Using the mean, we will compare with the deaths value to understand if the mean of the two samples are same. Here am performing the test of independence.

1.Test for Injury deaths raw value:

H0: The mean of Norm deaths and Injury deaths are the same

H1 : The mean of samples is different

Reject the Null Hypothesis as the pvalue is less than 0.05 The null hypothesis is rejected as the the mean of norm deaths and that of injury deaths are not same. Thus, we accept the alternative hypothesis.

2. Test for Drug overdose deaths raw value:

H0: The mean of Norm deaths and Drug overdose deaths raw value are the same.

H1: The mean of samples is different.

The null hypothesis is rejected as the the mean of norm deaths and that of drug overdose deaths raw value are not same. Thus we accept the alternative hypothesis

3. Test for Premature death raw value value:

H0: The mean of Norm deaths and Premature death raw value are same.

H1: The mean of samples is different.

The null hypothesis is rejected as the mean of norm deaths and that of Premature death raw value are not same. Thus we accept the alternative hypothesis.

4. Test for Insufficient sleep raw value:

H0: The mean of Norm deaths and Insufficient sleep raw value are same.

H1: The mean of samples is different.

The null hypothesis is rejected as the mean of norm deaths and that of Insufficient sleep raw value are not same. Thus, we accept the alternative hypothesis.

5. Test for Firearm fatalities raw value:

H0: The mean of Norm deaths and Firearm fatalities raw value are same.

H1: The mean of samples is different.

The null hypothesis is accepted as the mean of norm deaths and that of Firearm fatalities raw value are. Thus, we accept the null hypothesis.

**Linear Regression:**

Inference:

The independent variable is opioid dispensing rate, and the dependent variable is norm deaths.

The R-squared value is 0.001 i.e with just one independent variable — accounts for roughly 1% of the variance in death rates

we see most of the deaths were when the opioid dispensing rates was equal to 100 units.

From the above parameters we see that the slope is positive this means that with the increase in opioid dispensing rate, the deaths also increased.

We can interpret the coefficient of opioid dispensing rate as, every year the deaths for opioid dispensing rate is found to increase by 12%

**Multiple Linear Regression:**

Inference:

* Here the opioid dispensing rate is the dependent variable, and the rest other variables are independent.
* Injury\_deaths\_raw\_value - 0.093788, Adult\_smoking\_raw\_value - 26.258083, Firearm\_fatalities\_raw\_value - 0.051801 are positively related as there is a positive slope so with increase in one variable the dependent variable will also increase.
* Drug\_overdose\_deaths\_raw\_value and In-sufficient sleep are not related to opioid dispensing as the values are negative.
* The R-squared is 0.164 for this model and with 5 independent variables — accounts for roughly 16% of the variance in opioid dispensing rates.

**Polynomial linear regression degree – 2 , 3 , 4**

**Degree = 2**

**Inference:**

From the summary table we conclude the below:

R2 is the coefficient of determination that tells us that how much percentage variation dependent variable can be explained by independent variable. Here, 32.5 % variation in Opiod dispemsing rate can be explained by the 5 variables that we have chosen. The maximum R-squared value can be 1, which means that the larger the R-squared value better is the regression.

The coefficient term tells the change in opioid dispending rate for a unit change in independent variables i.e if independent variables rises by 1 unit then opioid dispending rate rises by 27% for drug overdose, 43% for firearm fatalities , 38% injury deaths, 2.6% in adult smoking and 1.27% decrease in insufficient sleep.

**Degree = 3**

**Inference¶**

* R2 is the coefficient of determination that tells us that how much percentage variation dependent variable can be explained by independent variable. Here, 33.1 % variation in Opioid dispensing rate can be explained by the 5 variables that we have chosen. The maximum R-squared value can be 1, which means that the larger the R-squared value better is the regression.
* The coefficient term tells the change in opioid dispending rate for a unit change in independent variables i.e if independent variables rise by 1 unit, then opioid dispending rate rises by 24.14% for drug overdose, 0.17% for firearm fatalities, 13.8% injury deaths, approx. 0.1% in adult smoking and approx. 0.1% in insufficient sleep.

**Degree = 4**

**Inference**

* R2 is the coefficient of determination that tells us that how much percentage variation dependent variable can be explained by independent variable. Here, 33.9 % variation in Opioid dispensing rate can be explained by the 5 variables that we have chosen. The maximum R-squared value can be 1, which means that the larger the R-squared value better is the regression.
* The coefficient term tells the change in opioid dispending rate for a unit change in independent variables i.e if independent variables rise by 1 unit, then opioid dispending rate rises by 24.14% for drug overdose, 63% for firearm fatalities, 75% injury deaths, approx. 0% in adult smoking i.e dependent and independent variables are not related and approx. 0 % in insufficient sleep as the dependent and independent variables are not related