**IAC 621 Data Science Stage 3 Member Task Report**

**Kyle Killworth**

**Part 1- Fitted State Distributions:**

After re-running code from Stage 2 to generate new weekly COVID cases and deaths for the states of Indiana, Ohio, Illinois, Michigan, Kentucky, and Wisconsin, the major step from Stage 3 was to plot the distribution of COVID cases for Indiana. This was done through both a histogram and a density plot. Ultimately density plots will be used for other states, since the COVID cases are per 100,000 people, and as such are a continuous variable.

Figure 1: Indiana Histogram

Chart, histogram

Description automatically generated

Figure 2: Indiana Density Plot

A picture containing shape

Description automatically generated

Based on the density plot above, Indiana appears to follow a gamma distribution for its new weekly COVID cases. What are its statistics?

Table 1: Descriptive Statistics

|  |  |
| --- | --- |
| Mean | 33.09345794392523 |
| Variance | 1663.1043907600067 |
| Skewness | 2.359975746241423 |
| Kurtosis | 6.336053135573369 |

The mean looks to be just over 33, and variance is nearly 1,664. Our skewness is quite high, but that's not surprising considering what we saw from the density plot above. Like with skewness, kurtosis is high, again not surprising considering the density plot. The next step is to generate the density plots for the other 5 states.

For details on the density plots, look at the related jupyter notebook, but the distributions are nearly identical for every state. This would indicate that each of the states have similar patterns in regard to weekly average covid cases. This makes sense, since all these states are geographically close to each other, so their COVID case patterns would be influenced by each other. We now are moving on to poisson distributions for each state. We'll use the mean value for Indiana cases as our parameter for generating our poisson distribution. In this example we chose to have a sample size of 100,000, since that is what our cases are multiplied by.

Figure 3: Indiana Cases Poisson Distribution

Chart, histogram

Description automatically generated

Unlike what we saw with our density plots, the poisson plot is much more symmetrical. This is in part due to large sample size, and also due to the nature of poisson distributions. They model the probability of seeing a certain number of successes within a time interval (in this case, one week). Its single parameter is a mean value, which as mentioned before is the mean number of cases per week. So, when sampling, it will be sampling around the mean, so values closest to the mean will be higher. If the average number of cases (successes) is around 33, then the distribution will peak around 33. The minimum (11) and maximum (60) values for this distribution have probability mass functions (pmf) so low that they go 6 or 7 decimal places into scientific notation. There is essentially a 7% chance of seeing exactly our mean of 33 cases per 100,000 in a week. This seems likely, as even though the mean is near 33, there are a lot of numbers that can be selected, so the odds of exactly 33 will still not be particularly high.

Not surprisingly, the other 5 states have distributions near identical to Indiana. The main difference comes from the mean point of the poisson distribution. Illinois is the most similar to Indiana, at around 32. Wisconsin looks to be at 33 as well, but numbers above it don't drop as quickly. Both Michigan and Ohio hover around 30. Interestingly, Kentucky has higher mean near 36 or 37. This data further confirms my previous statement that the 5 states closely mirror each other in Covid cases.

Figure 4: Indiana Deaths Poisson Distribution

Chart, histogram

Description automatically generated

This is where my prior point about higher mean parameters for a poisson distribution make sense. Since very few people per 100,000 die per week, the mean is very low, so the data becomes skewed. The chance of exactly 0 people per 100,000 is approximately 70%. Like we’ve seen before, this is essentially mirrored in the other 5 states. We can now move on to Part 2, looking at the Enrichment Data and how that correlates with COVID cases and deaths.

**Part 2- Enrichment Data Correlations:**

After pulling in the variables I am concerned with within the ACS Social and Economic Datasets and normalizing the various variables per 100,000 people, I ran correlations against the COVID cases:

Table 2: Enrichment Data Correlations with COVID Cases

|  |  |
| --- | --- |
| Number of Residents with No Diploma | 0.61 |
| Number of Residents with High School Diploma | 0.66 |
| Number of Residents with Associate Degree | 0.71 |
| Number of Residents with Bachelor’s Degree | 0.69 |
| Number of Residents with Graduate Degree | 0.64 |
| Number of Foreign-Born Residents | 0.53 |
| Number of Residents with less than 10k income | 0.59 |
| Number of Residents with 10-15k income | 0.65 |
| Number of Residents with 15-25k income | 0.65 |
| Number of Residents with 25-35k income | 0.64 |
| Number of Residents with 35-50k income | 0.68 |
| Number of Residents with 50-75k income | 0.67 |
| Number of Residents with 75-100k income | 0.68 |
| Number of Residents with 100-150k income | 0.70 |
| Number of Residents with 150-200k income | 0.67 |
| Number of Residents with more than 200k income | 0.60 |
| Number of Residents Below Poverty Level | 0.68 |

There's a lot to look at, but it's unfortunately not that useful. Pretty much every one of these variables has a correlation between 0.6 and 0.7. It’s unfortunate to see that none of the salary ranges provide any meaningful differences from one another. Basically, as each variable increases, so do COVID cases in a similar manner. This could potentially be due to how everything is connected to population. More population, more cases. Checking for deaths, the Foreign-Born variable does show a bit of a difference at 0.36. This is the lowest correlation of all the variables with COVID. While still positive, it does indicate that higher number of foreign-born residents does not have that much of an impact on COVID deaths. I did notice that there were some differences within the education variables and COVID deaths as well:

Table 3: Education Data Correlations with COVID Cases

|  |  |
| --- | --- |
| Number of Residents with No Diploma | 0.61 |
| Number of Residents with High School Diploma | 0.58 |
| Number of Residents with Associate Degree | 0.51 |
| Number of Residents with Bachelor’s Degree | 0.48 |
| Number of Residents with Graduate Degree | 0.51 |

While the correlations are still relatively strongly positive, the strengths definitely decrease as the degree gets more advanced (though graduates go back up a bit). This could be a signal that something notable can be found here. I think next steps would be to focus on a few variables and expand the number of states. Adding more states could lead to more relevant and accurate relationships (or lack thereof) between variables and COVID.

Hypothesis 1: Does an increase in foreign born residents lead to higher counts of COVID cases or deaths?

Hypothesis 2: Does having a higher number of residents with no diploma lead to higher counts of COVID cases or deaths?

Hypothesis 2: Does having a higher number of residents with a Bachelor's degree lead to higher counts of COVID cases or deaths?