Adithyan Subramanian

Inferential Statistics report

Mentor: Mukesh Mithrakumar

The overarching goal of this project is to recommend actions to the City of Chicago to promote healthiness and health equity within all neighborhoods. In addition, we will attempt to model and predict the lung cancer and infant mortality rates for neighborhoods in Chicago. In order to do this we must first perform some inferential statistical analysis. To do so I chose to perform a frequentist inference study.

This study aimed to answer two questions.

1) *Is there a significant difference in lung cancer rates in neighborhoods with high number of tobacco sellers compared to neighborhoods with a low number of tobacco sellers?*

2) *Is there a significant difference in infant mortality rates in neighborhoods with a high number of child related businesses compared to neighborhoods with a low number of child related businesses?*

The null hypothesis for the first question is that there is no difference in lung cancer rates in neighborhoods with a high number of tobacco sellers compared to neighborhoods with a low number of tobacco sellers. To run my frequentist test, I split the dataset so that any neighborhood with a lower than average number of tobacco sellers was considered "low tobacco neighborhood." any neighborhood with an above average number of tobacco sellers was considered a "high tobacco neighborhood." I then performed a 2-sided T-test. With the following results.

Ttest_indResult(statistic=-2.6192258278255816, pvalue=0.011372935200506825)

Given that the statistic is less than the p-value, and the p-value is less than 0.05, we reject the null hypothesis, meaning that there is a significant difference in lung cancer

rates in neighborhoods with a high number of tobacco sellers compared to neighborhoods with low number of tobacco sellers.

The null hypothesis for the second question is that there is no difference in infant mortality rates in neighborhoods with a high number of child related businesses compared to neighborhoods with a low number of child related businesses. Similarly to the first analysis, I split my dataset based on if a neighborhood had an above average number of child related businesses, or if a neighborhoods had a below average number of child related businesses. I again did a 2-sided T-test, with the following results.

Ttest_indResult(statistic=-0.06516825791640102, pvalue=0.9482522439304938)

Given that the p-value is greater than 0.05, we accept our null hypothesis. There is no significant difference in infant mortality rate based on the number of child related businesses.

Given the results of these two frequentist analysis, we can safely say that lung cancer rates are affected by the number of tobacco sellers within a neighborhood, and that infant mortality rates are not affected by the number of child-related businesses. Further frequentist testing can be used to analyze other businesses affect on various health metrics, but is beyond the scope of this project.