Viva Wan, Yaohan Xu

CPLN5050 Planning by Numbers

February 14, 2024

Assignment #1: Comparing Means, Association, and Correlation

Technical Addendum

# Task 1.

We firstly calculated the poverty rate by dividing the total population by the population in poverty, and then created a new variable named “type” to classify the Census tracts. We use “ifelse” function in R to categorize Census tracts that have travel time and poverty rate both below or equal to the median values as DTDs and others as DTOs. In such a case, we include the lower cutting value (median values) in the 3rd quartile.

The result shows that 112 census tracts are DTDs and 318 are DTOs in the Columbus metropolitan area.

# Task 2.

We firstly transformed the coordinate reference system of the driving schools’ shapefile into that of the census tracts’ shapefile, and then used ggplot to create the map as below. For the map, we used red for DTDs, green for DTOs and white dot with grey outlines for driving schools.

The map shows that DTOs are census tracts with driving schools either inside or nearby, while DTDs are areas that are distant from driving schools. Moreover, most DTDs are located at the marginal corners of Columbus, Ohio metro area.

A map of a neighborhood

Description automatically generated

# Task 3.

To calculate the population density, we divided the total population by the geographic area of corresponding census tract in square miles (total\_population/area\_sq\_mi). To calculate the vehicle ownership rate, we first determining the share of households with zero vehicles by dividing the number of households with zero cars by the total households and then subtracted this share from 1 (1-zero\_veh\_hh/total\_households). To calculate the percentage of non-White residents, we firstly got the percentage of White residents by dividing the White population by the total population and then subtracted the share from 1 (1- whites/total\_population).

# Task 4.

We firstly transformed the census tracts’ shapefile in to the tibble form first by dropping the geometry column, grouped them by the categories, and then calculated the maximum, minimum, mean, and standard deviation of travel time, median household income, poverty rate, population density, vehicle ownership rate and percent of non-White residents. Eventually, we exported the data into csv and formatted them in Excel as below (Table 4-1).

Table 4-1 Summary table of six variables for DTDs and DTOs

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| type | variable | maximum | minimum | mean | standard deviation | unit |
| DTD | **travel time** | 34 | 8 | 12 | 5 | minute |
| **median household income** | 89,537 | 12,565 | 43,908 | 14,891 | $ |
| **poverty rate** | 68% | 11% | 26% | 14% | % |
| **population density** | 26,253 | 20 | 4,353 | 4,617 | persons per square miles |
| **vehicle ownership rate** | 100% | 48% | 87% | 10% | % |
| **percentage of non-White residents** | 89% | 1% | 35% | 29% | % |
| DTO | **travel time** | 27 | 2 | 7 | 4 | minute |
| **median household income** | 189,531 | 16,528 | 75,405 | 33,821 | $ |
| **poverty rate** | 56% | 0% | 11% | 10% | % |
| **population density** | 13,348 | 32 | 3,253 | 2,702 | persons per square miles |
| **vehicle ownership rate** | 100% | 12% | 94% | 8% | % |
| **percentage of non-White residents** | 90% | 0% | 24% | 23% | % |

# Task 5.

We chose population density and percentage of non-White residents as the two variables of interest. For each variable, we used ggplot to create the histogram for DTDs and DTOs respectively and then created the boxplot. In all plots, DTDs shows in red and DTOs shows in green as below.

1. Population density

The histogram for DTDs’ population density shows a right-skewed distribution, indicating that most DTDs have low population density below 5000 people per square miles, with fewer tracts having higher density. The presence of bars at the higher end suggests there are some areas with significantly higher population densities over 15,000 people per square miles, although these are less common. (Figure 5-1)

The histogram for DTOs’ population density also shows a right-skewed distribution but more gradual across different travel times with fewer occurrences of very high population density. Most tracts have density below 8000 people per square miles. (Figure 5-2)

In the boxplot about population density, DTDs have a slightly higher median and a wider range of population density than DTOs. Both have higher outliers, but outliers of DTDs have much higher densities. (Figure 5-3)

1. Percentage of non-White residents

The histogram for DTDs’ non-White population share shows a somewhat uniform distribution across the range, with a slight concentration at the lower end, indicating a number of DTDs with a very small percentage of non-White population around 5%. Several peaks at several intervals, suggesting that certain percentages of non-White population are more common in DTDs. such as 32.5%, 62.5% and 81%. (Figure 5-4)

The histogram for DTOs’ non-White population share shows a right-skewed distribution with a clear concentration of tracts with a low percentage of non-White population around 6-8%. The frequency steadily decreases as the percentage of non-White population increases, showing that higher percentages of non-White populations are less common in DTOs. (Figure 5-5)

In the boxplot for DTDs’ and DTOs’ non-White residents share, DTDs have a higher median and a wider range of non-White population percentage than DTOs. Only DTOs have outliers at the upper end, which means some DTOs have an unusually high percentage of non-White population, although they are not the norm. (Figure 5-6)

A graph of population density

Description automatically generated

Figure 5-1 Histogram for DTDs’ population density

A graph with green bars

Description automatically generated

Figure 5-2 Histogram for DTDs’ population density

A diagram of a graph

Description automatically generated with medium confidence

Figure 5-3 Boxplot for DTDs’ and DTOs’ population density

A graph of red lines

Description automatically generated with medium confidence

Figure 5-4 Histogram for DTDs’ non-White population share

A green and white graph

Description automatically generated

Figure 5-5 Histogram for DTOs’ non-White population share

A red and green squares

Description automatically generated

Figure 5-6 Boxplot for DTDs’ and DTOs’ non-White population share

# Task 6.

1. T-test on average probabilities of enrolling

DTDs have a statistically significant lower average probabilities of enrolling than DTOs (t = -9.25, p<0.05). DTDs and DTOs have the mean average probabilities of enrolling of 34% and 57% respectively, with a difference of 23%. Among all the census tracts, the difference was likely to be between 18% and 27%. (Figure 6-1)

1. T-test on vehicle ownership rates

DTDs have a statistically significant lower vehicle ownership rates than DTOs (t = -6.42, p<0.05). DTDs and DTOs have the mean vehicle ownership rates of 87% and 94% respectively, with a difference of 7%. Among all the census tracts, the difference was likely to be between 8% and 4%. (Figure 6-2)

1. T-test on non-White population share

DTDs have a statistically significant higher percentage of non-White population than DTOs (t = 3.7, p<0.05). DTDs and DTOs have the mean percentage of non-White population of 35% and 24% respectively, with a difference of 11%. Among all the census tracts, the difference was likely to be between 5% and 2%.

1. Assumptions violated

Firstly, those data violates the assumption of normal distribution. Some data distributions are right-skewed, including the percentage of non-White population for both DTDs and DTOs, and the average probabilities of enrolling for DTDs. Others are left-skewed, including the average probabilities of enrolling for DTOs and vehicle ownership rates for both. (Figure 6-4)

Secondly, those data also violates the assumption of independent observations that requires all the observations are unrelated. Among census tracts, neighboring tracts always correlate with each other in some ways.

A white text with black text

Description automatically generated

Figure 6-1 T-test on average probabilities of enrolling

A white text with black text

Description automatically generated

Figure 6-2 T-test on vehicle ownership rates

A white paper with black text

Description automatically generated

Figure 6-3 T-test on non-White population share

A graph of a number of probabilities

Description automatically generated A graph of green bars

Description automatically generated

A graph of a car ownership rate

Description automatically generated A graph with green and red bars

Description automatically generated

Figure 6-4 Histograms of variables

# Task 7.

We firstly calculated the cut-off values for the quartile of both median household income and the average probability of enrolling in driver training, and then created two variables to labels census tracts by their location in the quartiles. For the quartile, we included the lower cutting value in each quartile as we did for task 1. Afterwards transforming the data into tibble form, we calculated the number of each probability’s quartile that falls in each income’s quartile and then got the share of each bar in each income group. Afterwards, we exported the data into csv and made the group bar chart in Excel as below (Figure 7-1).

From the chart, we saw a clear pattern that higher median household income is associated with higher probabilities of enrolling, suggesting a positive correlation between income and enrollment probabilities. Census tracts in the lowest income quartile have the lowest probabilities of enrolling, with a steep decrease from the 1st to 4th quartile of enrolling probability within this group (78% to 1%). Those in the 2nd quartile have a noticeable increase in enrollment probabilities from the 1st to the 2nd quartile in the second income group, but the share of high probability (4th quartile) still remains significantly lower than that of higher income groups. In 3rd quartile income group, the share of low probability (1st quartile) decreases greatly (1%) while high probability groups grow in share and spread relatively evenly across the quartiles ranging from 21% to 48%. The highest income group has the highest probabilities of enrolling across all of its quartiles, with the 4th quartile at 74%, indicating a very high likelihood of enrollment compared to the other groups. In summary, the highest income households (4th quartile income) have a notably higher likelihood of enrolling (4th quatile enrolling = 78%) compared to lowest income households (1st quartile income, 4th quatile enrolling = 1%), with the probabilities increasing as income increases.

A graph of a number of people

Description automatically generated

Figure 7-1 Group bar chart

# Task 8.

We used CrossTable function to run Chi-squared test on the quartile groups of income and probabilities that we created from Task 7 (Figure 8-1). And then we used cramersV function to calculate the practical effect (Figure 8-2).

The result shows a significant association between income and probabilities of driver training enrollment (Chi^2 = 408, df = 9, p<0.05). Based on the standardized residuals, the association was primarily driven by more census tracts within the lowest income quartile group than expected that fall into the lowest probability quartile group of driving training enrollment (84 observed v.s. 27 expected, Std Residual = 10.9) as well as more census tracts within the highest income quartile group than expected that fall into the highest probability quartile group of driving training enrollment (80 observed v.s. 27 expected, Std Residual = 10.2). Moreover, not only the Chi-square is big, meaning significant association, the association is strong to moderate (cramersV = 0.56).

A paper with numbers and a few words

Description automatically generated with medium confidence

Figure 8-1 Chi-squared test

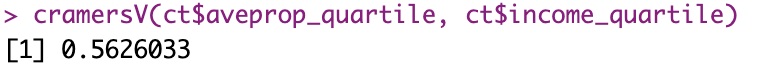


Figure 8-2 CramersV to calculate the effect size

# Task 9.

We firstly plotted the population density and the average probability of driving training enrollment (Figure 9-1). There is a high density of data points at the lower end of population density but at the higher end of average probability, which suggests that there are many census tracts with low population density and high average probability of driving training enrollment. There appears to be a negative relationship between the two variables, because as population density increases, the average probability seems to decrease. There are some outliers, particularly noticeable at the higher end of the population density and low end of average probability, meaning that a few census tracts with extremely high population density have very low average probability of driving training enrollment.

Then we plotted the non-White population share and the average probability of driving training enrollment (Figure 9-2). There is a large clustering of data points towards the lower end of the non-White Population Share and the higher end of the average probability, indicating that many census tracts predominated by white population have high average probability of driving training enrollment. There seems to be a slight negative correlation between the non-White population share and the average probability. As the share of the non-White population increases, the average probability tends to decrease.

A graph of a number of dots

Description automatically generated

Figure 9-1 Scatterplot of population density and probability of driving training enrollment

A graph of dots and numbers

Description automatically generated

Figure 9-2 Scatterplot of non-White population share and probability of driving training enrollment

# Task 10.

We firstly ran a correlation test on probability and two variables respectively and then calculated the coefficient of determination by squaring the Pearson’s r.

The correlation test between probability and population density shows that the average probability of driving training enrollment is negatively correlated with population density with a Pearson’s correlation coefficient of -0.59. This is a relatively large effect and indicates that 35% of the variability in average probability is shared by the population density. This correlation is statistically significant at p-value = 0.05 level (p-value < 0.05, N = 428). (Figure 10-1)

The correlation test between probability and non-White population share shows that the average probability of driving training enrollment is negatively correlated with non-White population share with a Pearson’s correlation coefficient of -0.72. This is a relatively large effect and indicates that 52% of the variability in average probability is shared by the non-White population share. This correlation is statistically significant at p-value = 0.05 level (p-value < 0.05, N = 428). (Figure 10-2)

A computer screen shot of a code

Description automatically generated

Figure 10-1 Correlation test between probability and population density

A screenshot of a computer

Description automatically generated

Figure 10-2 Correlation test between probability and non-White population share