Introduction

The USArrests datset is a dataset containing statistics related to violent crime in each of the 50 US states in 1973, it contains information on assault, murder and rape and the percentage of population residing in urban areas. The dataset has 50 observations (one for each state) and 4 variables: `Murder` (murder arrests per 100,000 residents), `Assault` (assault arrests per 100,000 residents), `UrbanPop` (percent urban population), and `Rape` (rape arrests per 100,000 residents).



Hypothesis 1

It is believed that with the increase in Urban population or the shift from rural to urban has made living conditions and employment opportunities worse in the urban area, leading to more and more crimes. People resorting to crime as a source of income to support their families in not uncommon. This hypothesis tests whether there is any relationship between the rate of UrbanPop and the murder rates.

Null Hypothesis (H0): There is no significant relationship between the rate of UrbanPop (urban population percentage) and the Murder rate across the states.

Alternative Hypothesis (H1): There is a significant relationship between the rate of UrbanPop and the Murder rate across the states.

We are going to perform a normal correlation analysis between the variables UrbanPop and Murder. And will attempt to determine any significant relationship between the two

Hypothesis 2

Small scale robberies and thefts have become extremely common in urban areas. Cases of people assaulting each other is a common sight. This hypothesis tests if there is any relation between the Assault rates in high and low UrbanPop states.

Null Hypothesis (H0): There is no significant difference in the Assault rates between states with high and low UrbanPop percentages.

Alternative Hypothesis (H1): States with high UrbanPop percentages have significantly higher Assault rates compared to states with low UrbanPop percentages.

Analysis: To test this hypothesis, we would first categorize the states into two groups based on UrbanPop percentages (e.g., high vs. low). Then, we will perform a hypothesis test, such as a t-test or a Wilcoxon test, to compare the Assault rates between the two groups of states and determine if there is a significant difference.

Methodology and Results

Methodology employed is very simple. A total of three tests are chosen which are to be performed on the data set.

We first check the correlation between the variables ‘UrbanPop’ and ‘Murder’. The correlation coefficient quantifies the strength and direction of the linear relationship between these two variables. A positive value indicates a positive association, while a negative value indicates a negative association. The `cor\_test` performs a statistical test to determine if the correlation is significantly different from zero. The output will show the correlation coefficient and the test result. The correlation coefficient between the two variables comes out to be approximately 0.07, which indicates a very weak positive relation. The p-value from the correlation test is 0.6312, which is greater than the common alpha level of 0.05. This means that the correlation is not statistically significant at the 5% level. We fail to reject the null hypothesis that the correlation is zero.

A screen shot of a computer

Description automatically generated

We use T test to check the next hypothesis of assault vs Population Type. We first classify the states into high or low urban pop by using a simple Boolean operator of more than (>). If the urban pop is more than the mean of total urban pop of the data set it becomes a high urbanpop state if not then it comes under low urbanpop state. It then performs a t-test to compare the `Assault` rates between these two groups. The t-test assesses whether the difference in `Assault` rates is statistically significant between high and low urban population states. . The t-value is 1.6136 and the p-value is 0.1138. Since the p-value is greater than 0.05, we fail to reject the null hypothesis that the means are equal. This means there is not a statistically significant difference in `Assault` rates between states with high and low `UrbanPop` percentages at the 5% level.

A screen shot of a computer

Description automatically generated

We also use F test to test the variability or spread of the variable ‘Assault’ if they differ significantly between the two groups of states. The F-test is used to compare the variances of `Assault` rates between states with high and low `UrbanPop` percentages. The F-value is 0.88622 and the p-value is 0.7578. Since the p-value is greater than 0.05, we fail to reject the null hypothesis that the variances are equal. This means there is not a statistically significant difference in the variances of `Assault` rates between states with high and low `UrbanPop` percentages at the 5% level.

A screen shot of a computer

Description automatically generated