Experiment 6

Hypothesis Testing

Instruction: Create a new .Rmd file and write code for each section in separate R code block

Hypothesis Testing checks if data provides enough evidence to reject a default claim. You assume the **null hypothesis** is true, compute a test statistic, obtain a **p-value**, and compare it to **α**. If p ≤ α, reject the null; otherwise, keep it. It's a controlled yes/no decision.

Type of Tests in this experiment:

A. **Z-test**: A z-test is valid whenever your test statistic is (approximately) normally distributed with known/estimable standard deviation.
B. **Chi-square test**: Chi-square test for independence checks whether two categorical variables are related. It compares observed counts to expected counts under independence; small p-value means variables are associated.

Dataset: The Boston Housing Dataset contains 506 rows and 14 columns. The description of columns is as follows: **crim** – Per-capita crime rate; **zn** – % of residential land zoned for large lots; **indus** – % of non-retail industrial area; **chas** – 1 if tract borders Charles River; else 0; **nox** – Nitric oxide concentration (pollution); **rm** – Average number of rooms per dwelling; **age** – % of owner-occupied units built before 1940; **dis** – Distance to Boston employment centres; **rad** – Index of highway accessibility; **tax** – Property tax rate per $10,000; **ptratio** – Student–teacher ratio; **black** – 1000(Bk – 0.63)^2; historic demographic index; **lstat** – % of lower-status population; **medv** – Median home value (target).

1. Download the Boston dataset from library MASS. Find the column wise mean and standard deviation values. These are the true population parameters for this experiment.

2. **One sample Mean Z-Test**: Consider the **rm** column. 40 samples are randomly selected. A Null hypothesis is given to you that that true mean (u0) is equal to 6.2. Based on the available samples, suggest whether you should accept this hypothesis or reject this hypothesis. Assume a significance level of 5%.

3. **One sample proportion Z-Test**: Consider the **crim** column. Create an indicator variable (0 or 1) high_crime. A location is high crime if the crime rate is greater than median crime rate. We are interested in the proportion of locations which have a high crime rate. 40 samples (size=40) are randomly selected. A Null hypothesis is given to you that that true proportion (p0) is equal to 0.5. Based on the available samples, suggest whether you should accept this hypothesis or reject this hypothesis. Assume that the standard deviation of sample proportion is given by sqrt(p0*(1-p0)/size). Assume a significance level of 5%.

4. **Two-Sample Difference of Means Z-Test**: Consider the **medv** column. Split the column into two sets: locations where **chas**=1 and locations where **chas**=0. 20 samples are randomly selected from each set. A Null hypothesis is given to you that that true mean of both sets are equal. Based on the available samples, suggest whether you should accept this hypothesis or reject this hypothesis. Assume that standard deviation of difference of two sample means is given as total_sd=sqrt(sam1_sd^2+sam2_sd^2), where sam1_sd and sam2_sd are the standard deviation of individual sample means. Assume a significance level of 5%.

5. **Two-Sample Difference of proportions Z-Test**: Consider the **tax** column. Create an indicator variable (0 or 1) high_tax. A location is high tax if the tax is greater than median tax. We are interested in the proportion of locations which have a high tax rate. Split the column into two sets: locations where **chas**=1 and locations where **chas**=0. 20 samples are randomly selected from each set. A Null hypothesis is given to you that that true proportions of both sets are equal. Based on the available samples, suggest whether you should accept this hypothesis or reject this hypothesis. Assume that standard deviation of sample proportion is sqrt(p(1-p)(2/size)) where p is the combined proportion of both sets. Assume a significance level of 5%.

6. **Correlation Z-Test**: Consider the **rm** and **medv** columns. Find the correlation using all samples. This is the true correlation between these two variables. Sample 40 observations randomly and test the Null Hypothesis that the two variables are uncorrelated. Use Fisher Transformation 0.5*log((1+r)/(1-r)) to convert r to normally distributed random variable. The standard deviation of the obtained random variable is sqrt(1/(size-3)). Assume a significance level of 5%.

7. **Chi-square Independence Test**: Create two indicator variables as high_crime and high_tax. Run chi-square independence test on these two variables and identify if these variables are independent. A chi-square test assumes a Null Hypothesis that two variables are independent. Assume a significance level of 5%.

8. **Exploratory**: Repeat step 2 for various Null Hypothesis: u0={5, 5.6, 6.2, 6.8, 7.4}. Record your p-values in each case.

9. **Exploratory**: Repeat step 3 for various Null Hypothesis: p0={0.2,0.4,0.6,0.8}. Record your p-values in each case.