Case Study : US Health Insurance

Problem Statement

In this project, we attempt to analyze and explore the US Health Insurance dataset for medical costs in order to derive valuable insights, and find answers to questions through statistical hypothesis testing.

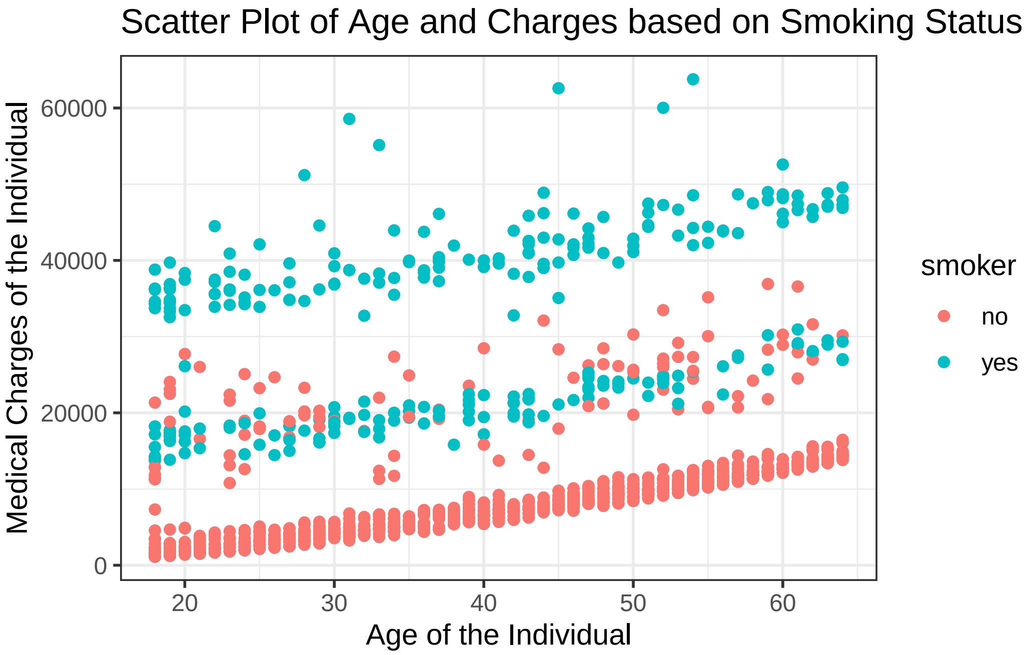
Objectives: 1. Exploratory Data Analysis 2. Hypothesis Testing

The data contains medical costs of people characterized by certain attributes. Viewing the summary of the data.

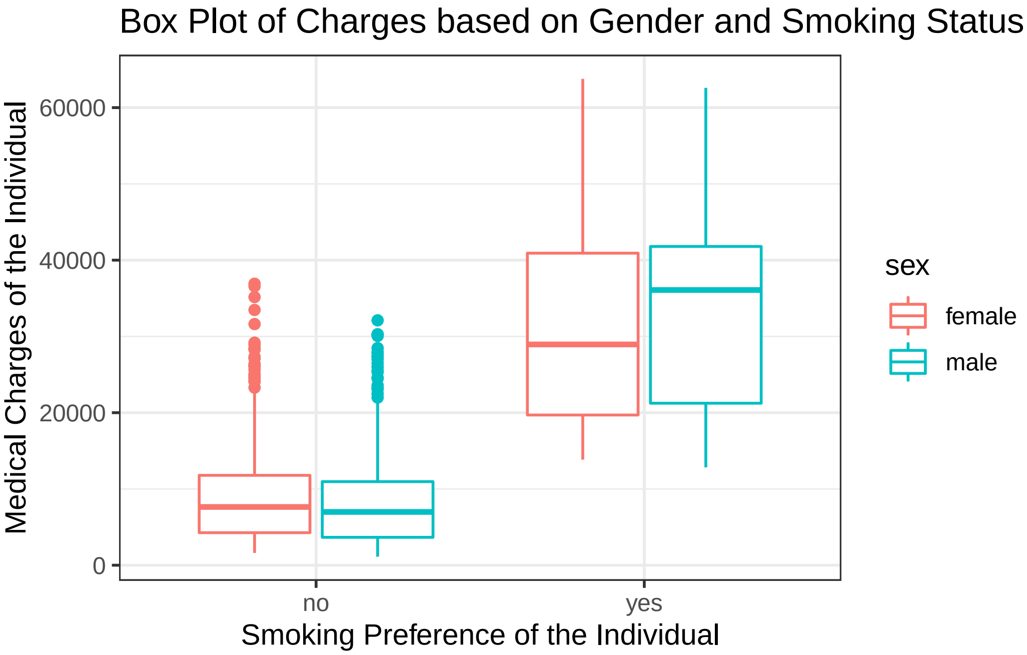
age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622

1. Graphical Analysis

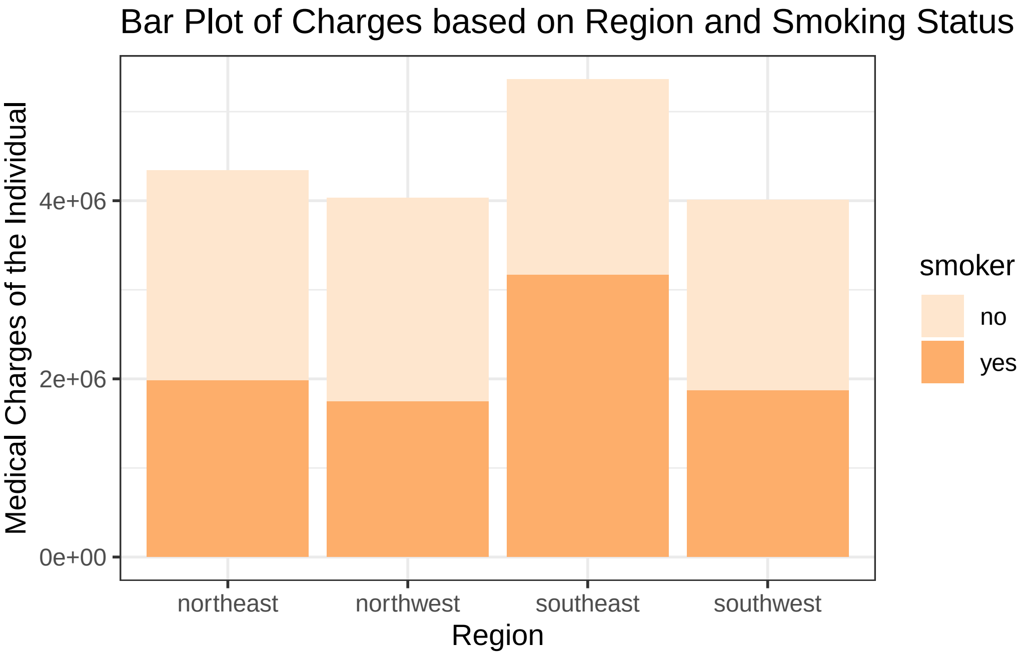
The relationship between insurance charges, age and smoking status (smoker/non-smoker) can be seen using a scatter plot. As can be observed, with increasing age the insurance charges increase in individuals. In addition, the insurance charges are higher for individuals who smoke.



Exploring the relationship between insurance charges, gender, and smoking status (smoker/non-smoker) using a box plot. As can be observed, there is a significant difference in the insurance charges based on smoking status. Individuals who smoke have a higher insurance charge than individuals who do not. Moreover, even for individuals who smoke, the insurance chargers are higher in men.



Viewing the changing in the insurance charges based on the region and smoking status using a bar plot. The medical charges of individuals are higher in the SouthEast region. Even so, the number of smokers in this region are higher compared to other regions. Are the number of smoking individuals increasing the overall medical charges in the SouthEast Region?



2. Joint Distribution for Age and Number of Children

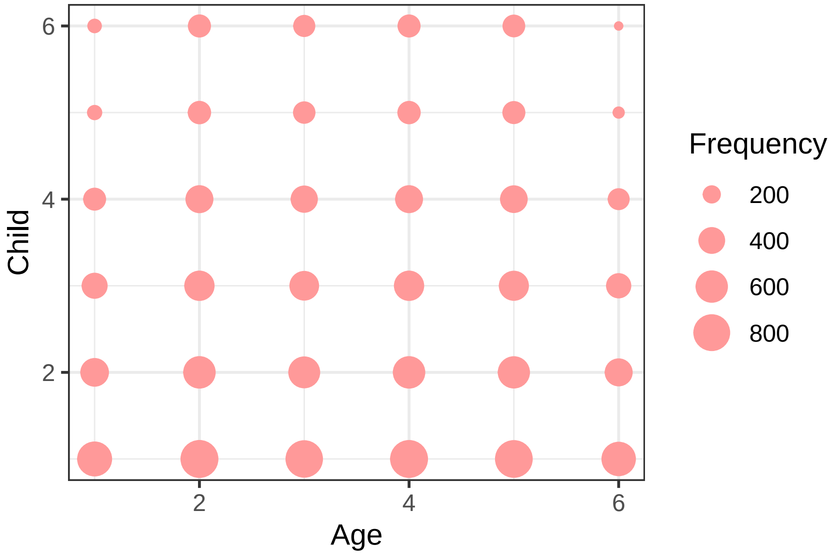
To get an idea on how frequently an individual can be of a particular age group and have children (either no children or have children), we obtain the joint probability distribution.

The joint probability distribution is obtained as follows:

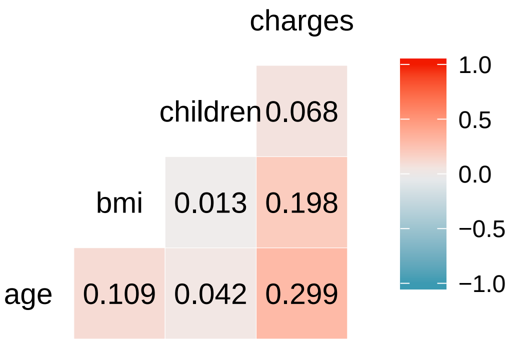
```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
## [1,] 0.044 0.029 0.023 0.018 0.010 0.010
## [2,] 0.053 0.038 0.032 0.027 0.019 0.019
## [3,] 0.052 0.036 0.031 0.026 0.018 0.017
## [4,] 0.053 0.038 0.032 0.027 0.019 0.018
## [5,] 0.053 0.037 0.032 0.027 0.018 0.018
## [6,] 0.043 0.027 0.022 0.017 0.009 0.008
```

3. Are age and number of children correlated?

Visual representation of the frequency of age and number of children.



All combinations of age and children are **weakly correlated**. The correlation coefficient of age and children is 0.04.



4. Does the average charge of individuals with children differ from that without children in the Northeast region?

Step 1 : Define the Hypothesis

X1 = RV of charges of individuals without children in the northeast region X2 = RV of charges of individuals with children in the northeast region

Null hypothesis H0 : mu1 = mu2 Alternative hypothesis H1 : mu1 != mu2

Step 2 : Calculate the Test Statistic

The calculated value of Zcal is -2.649.

Step 3 : Decision Rule

The upper and lower bound of the acceptable region is -1.96 and 1.96. The significance level is 0.05. The value of Zcal -2.649 is in the rejection region. Hence, we reject the null hypothesis.

Step 4 : P Value

This is a two-tailed test. Therefore, p-value = 2 * p(z > Zcal).

The p-value is 0.008. P-value is less than alpha (0.05) so we can reject the null hypothesis.

Conclusion:

We Reject the Null Hypothesis and state that at 5% significance level, the mean average charge of individuals with and without children in the Northeast region are NOT EQUAL.

Hence, the mean average charge of individuals with and without children differ in the Northeast region.

5. Does the ratio of individuals who smoke in the Northeast Region differ from that in the Southwest region?

Step 1 : Define the Hypothesis

X1 = RV of individuals who smoke in the northeast region X2 = RV of individuals who smoke in the southwest region

Null hypothesis H0 : p1 = p2 Alternative hypothesis H1 : p1 != p2

p1 = Smokers / Total Individuals in the Northeast Region p2 = Smokers / Total Individuals in the Southwest Region

Step 2 : Calculate the Test Statistic

The calculated value of Zcal is 0.915.

Step 3 : Decision Rule

The upper and lower bound of the acceptable region is -1.96 and 1.96. The significance level is 0.05. The value of Zcal 0.915 is in the acceptable region. Hence, we fail to reject the null hypothesis.

Step 4 : P Value

This is a two-tailed test. Therefore, p-value = 2 * p(z > Zcal).

The p-value is 0.36. P-value is greater than alpha (0.05) so we fail to reject the null hypothesis.

Conclusion:

We Fail to Reject the Null Hypothesis and state that at 5% significance level, the ratio of individuals who smoke in the Northeast Region and the Southeast region are EQUAL. We can also view this graphical in the bar plot on page 3.

Hence, the ratio of individuals who smoke in the Northeast Region do not differ from that in the Southeast region.

6. Is the ratio of variances across all regions different for individuals above and below the age of 50 years?

Step 1 : Define the Hypothesis

X1 = RV of individuals above and equal to the age of 50 X2 = RV of individuals below the age of 50

Null hypothesis H0 : sigma1 = sigma2 Alternative hypothesis H1 : sigma1 != sigma2

Step 2 : Calculate the Test Statistic

The calculated value of Fcal is 1

Step 3 : Decision Rule

The upper and lower bound of the acceptable region is 0.8429961 and 1.1790763. The significance level is 0.05. The value of Fcal 1 is in the acceptable region. Hence, we fail to reject the null hypothesis.

Conclusion:

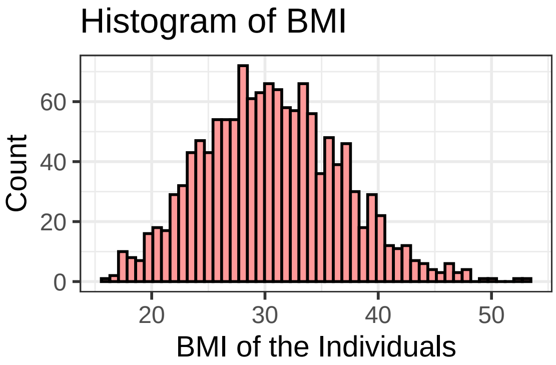
We Fail to Reject the Null Hypothesis and state that at 5% significance level, the ratio of variances across all regions different for individuals above and below the age of 50 years are EQUAL.

Hence, the ratio of variances of individuals above the age of 50 years do not differ from that below the age of 50 years.

7. Distribution Fitting

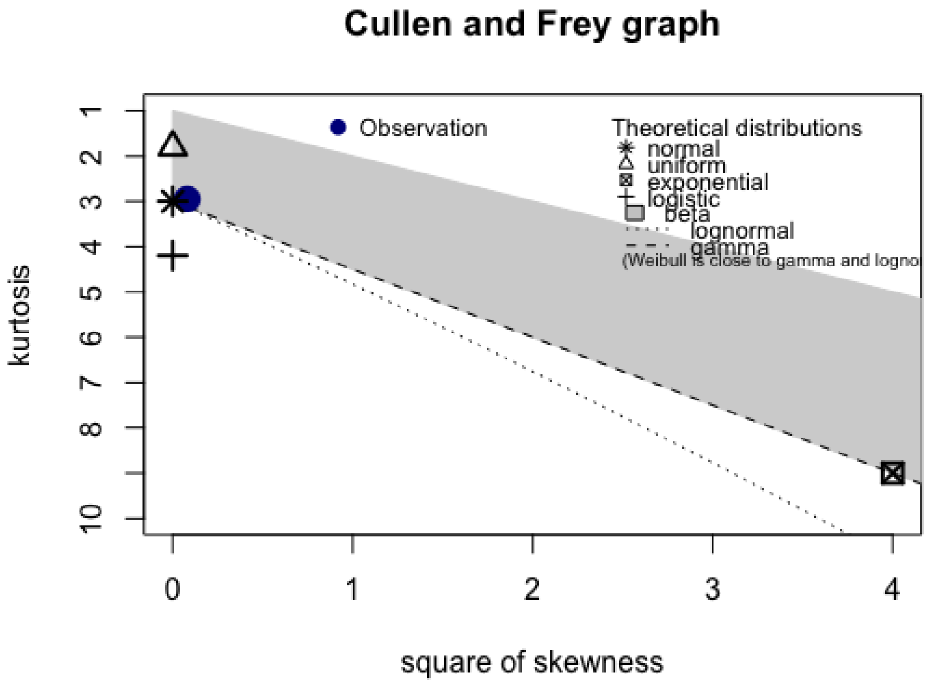
Visualization for BMI

Before fitting any distribution to a dataset, we will visualize the data in order to get an idea of what distributions are more likely to fit the data as compared to others.



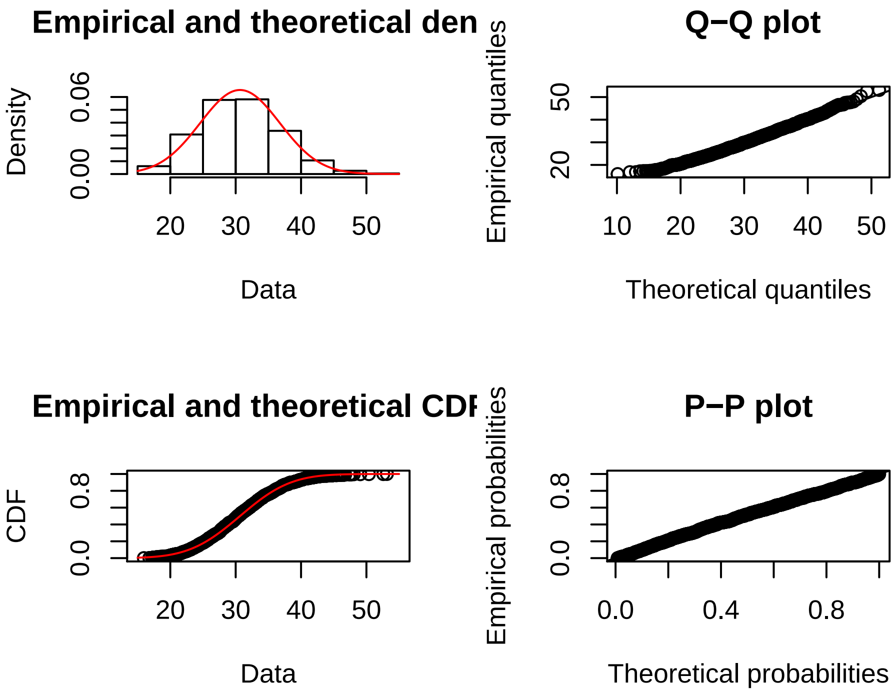
From observation, it seems that BMI has a normal like distribution.

Descriptive Statistics



The estimated kurtosis is 2.95. For the normal distribution, this value should be equal to 3. The estimated skewness is 0.28. Thus, the data is slightly positively skewed.
Fitting a normal distribution to the data.

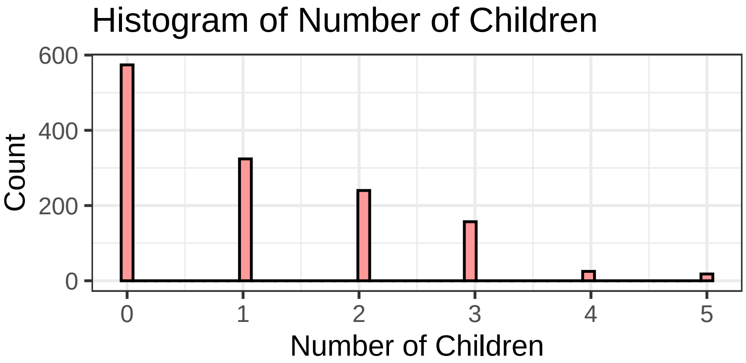
Goodness-of-fit Plots



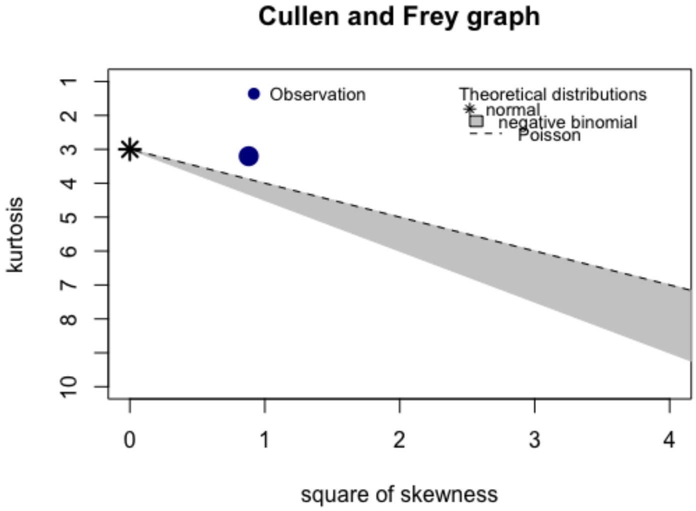
The AIC and BIC values for the distribution are 8638.264 and 8648.662, respectively.

Visualization for Number of Children

Visualizing the data in order to get an idea of what distributions are more likely to fit the data as compared to others.



Descriptive Statistics



From observation, it seems that the number of children could have a poisson or negative binomial distribution.

Fitting both distributions, to see which distribution gives better results.

For the poisson distribution, the AIC and BIC values for the distribution are 3892.885 and 3898.084, respectively.
For the negative binomial distribution, the AIC and BIC values for the distribution are 3825.728 and 3836.126, respectively. From this test, we see that both negative binomial and poisson distribution fit the data well. However, the negative binomial values are slightly better.

8. Is the proportion of individuals who smoke different across different genders?

Step 1 : Define the Hypothesis

X1 = RV of male individuals who smoke X2 = RV of female individuals who smoke

Null hypothesis H0 : p1 = p2 Alternative hypothesis H1 : p1 != p2

p1 = Male Smokers / Total Male Individuals p2 = Female Smokers / Total Female Individuals

Step 2 : Calculate the Test Statistic

The calculated value of Zcal is 2.787.

Step 3 : Decision Rule

The upper and lower bound of the acceptable region is -1.96 and 1.96. The value of Zcal 2.787 is in the rejection region. Hence, we can reject the null hypothesis.

Step 4 : P Value

This is a two-tailed test. Therefore, p-value = 2 * p(z > Zcal).

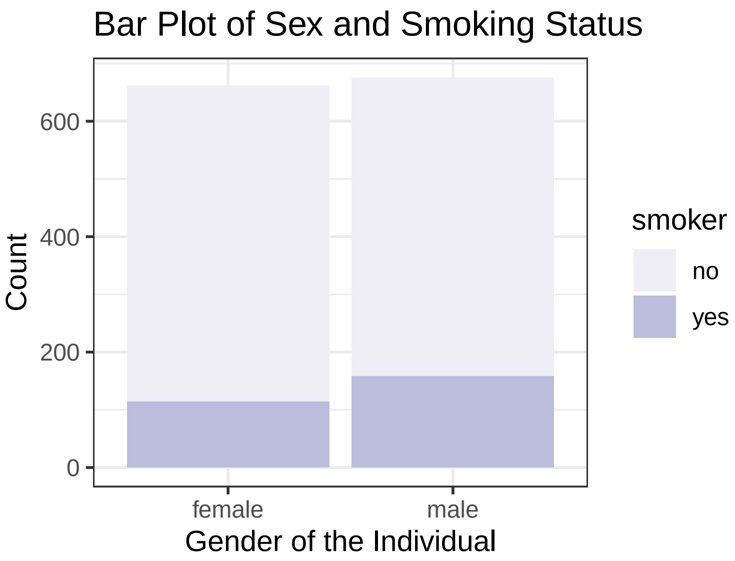
The p-value is 0.005. P-value is less than alpha (0.05) so we reject the null hypothesis.

Conclusion:

We Reject the Null Hypothesis and state that at 5% significance level, the proportion of individuals who smoke across different genders is NOT EQUAL.

Hence, the proportion of male individuals who smoke is different from that of female individuals who smoke.

This can also be observed using a bar plot as shown below.


Bar Plot of Sex and Smoking Status
Count
female
male
Gender of the Individual
smoker
no
yes