**Introduction**:

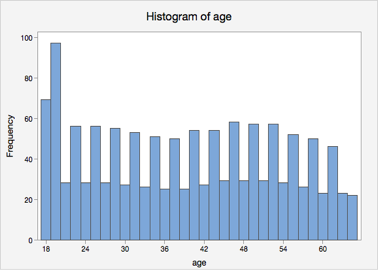
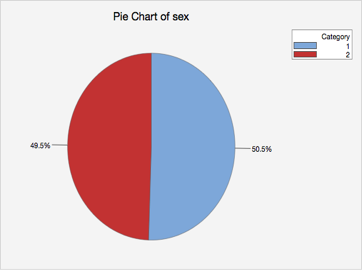
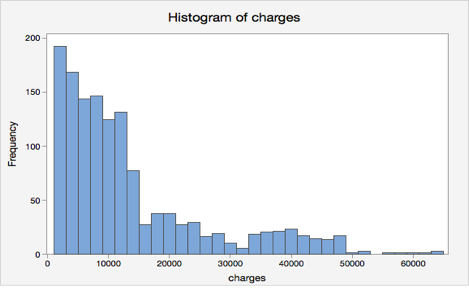
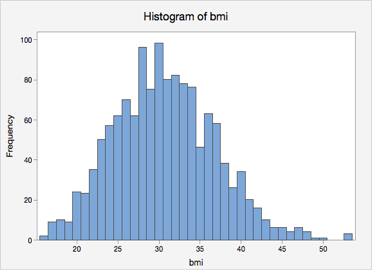
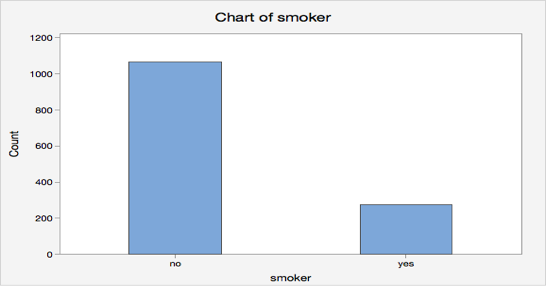
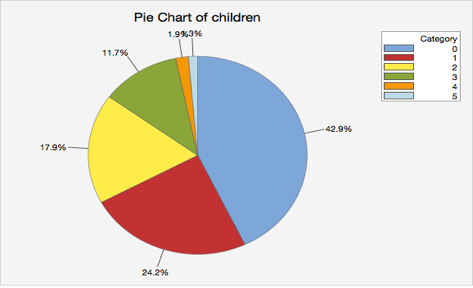
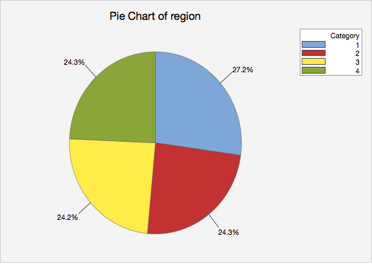
Health insurance covers the medical expenses for individual’s illness and injuries. The insurance charge depends up on various factors for each individual. A sample dataset is being analyzed and I have come up with a model to predict the approximate insurance charge for an individual in USA.

**Data**:

The data we have is a sample dataset obtained from different regions in equal proportions. We can call this type of sampling a **stratified random sampling**, as we have obtained equal proportions of data from different region to conclude model for the entire population. We have factors or variables like Age,Sex,BMI,Children a person have, region ,having smoking history and finally the insurance Charges that each person pays.

**Summarizing the Data:**

The Age data were uniformly distributed. The gender ratio was almost equal. BMI for each person was normally distributed. Ratio of People having dependent to people without dependent was 40%-60%. The average ratio for the four regions was 25%. In the dataset about 80% were non-smokers. The insurance charges were right skewed.



**Hypothesis testing:**

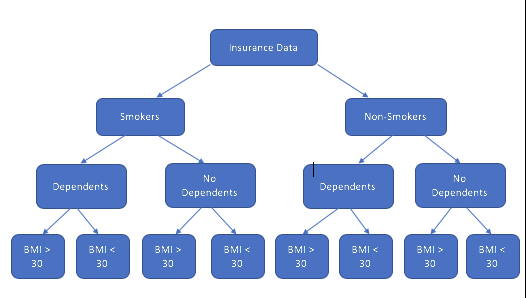
We assume that all variables have an influence on the insurance charges. Now, we have to prove that the charges is not dependent on all. So, our Hypothesis are as follows:

H0: β1=β2=β3=β4=β5=β6=0 (All of the variables have an influence on Insurance charges)

H1: Either of the β ≠ 0 (At least one of the variables doesn’t have an influence on Insurance charges).

I Performed a multiple regression on them with 95% confidence level. p value of Sex and Region was greater than 0.05 so we **reject Null hypothesis** and say not all the variables influence the insurance rates.

**Creating a model to predict Insurance rates - Regression:**

****We are splitting the Dataset into different modules as below:

On performing regression on each sample data in excel, we got the regression equation for each to predict the Insurance rates.

**R Snippet:**

Insurance\_Rate <- function(AGE, BMI, CHILDREN, SMOKER){

if(SMOKER == "yes" && CHILDREN == 0 && BMI < 30){

PRICE = (251.20 \* AGE) + (505.18 \* BMI) -956.74

} else if(SMOKER == "yes" && CHILDREN == 0 && BMI >= 30) {

PRICE = (292.16 \* AGE) + (614.01 \* BMI) + 8120.10

} else if(SMOKER == "yes" && CHILDREN > 0 && BMI < 30){

PRICE = (259.48 \* AGE) + (359.27 \* BMI) + 2428.48

} else if(SMOKER == "yes" && CHILDREN > 0 && BMI >= 30){

PRICE = (253.72 \* AGE) + (447.91 \* BMI) + 16021.03

} else if(SMOKER == "no" && CHILDREN == 0 && BMI < 30){

PRICE = (277.56 \* AGE) + (22 \* BMI) -3791

} else if(SMOKER == "no" && CHILDREN == 0 && BMI >= 30){

PRICE = (254 \* AGE) + (-68.2 \* BMI) -186.83

} else if(SMOKER == "no" && CHILDREN > 0 && BMI < 30){

PRICE = (249 \*AGE) + (178 \* BMI) -5204.2

} else {

PRICE = (252 \* AGE) + (165 \* BMI) -6421.75

}

return(PRICE)

}

**Sample Output:**

On giving input as - Insurance\_Rate(40, 25, 2, "NO")

The output value – 7783.25

When we give a person’s age, bmi, no.of children and give yes or no for smoker we get the approximate insurance charge that a person needs to pay.

**Conclusion:**

Using the sample dataset obtained from different regions, we were able to identify the significant variables that affect the insurance rate using hypothesis testing. We later formulated equations and wrote a R code snippet which could be useful to predict the insurance rates for an individual given, the respective variable values.