CASE STUDY : A Deep-Dive into The Health Insurance in The United States of America

MEERA NAGARIA (NU ID: 001023375)

4/18/2020

OBJECTIVE

The following case study helps us to enhance our analytical abilities by applying various concepts learnt in the Probability and Statistics class to real time dataset of US Healthcare industry.The statistical analysis has been performed using an open source programming language called R as it is popular for its graphical capabilities as well.

Description of US Health Insurance Dataset

There are 7 different entities like *Age, Sex, BMI, Number of Children, Smoker, and Region* that determine the *Charges* for the health insurance. To perform deep-dive analysis, the data was sliced and diced into smaller meaningful chunks further giving rise to 2 more entities/columns viz. Age-group [(0,9),(10,19),..(60,69)], Meets_BMI[“underweight”(<18.5),“Normal” (18.5 to 24.9),“overweight”(>25)].

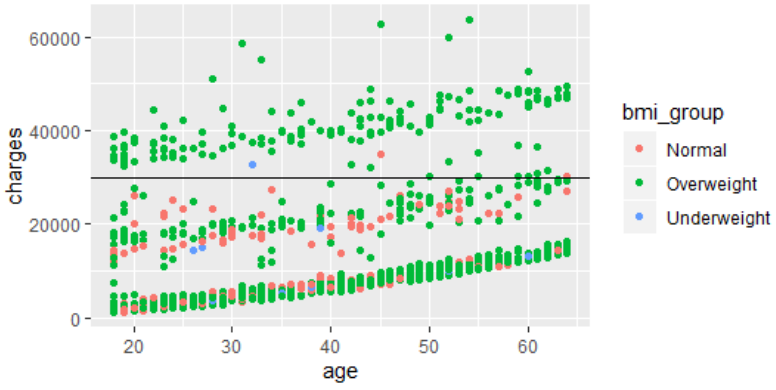
Problem Statement 1:

Finding meaningful insights from the dataset by performing graphical analysis.

Output 1: Graphical Analysis

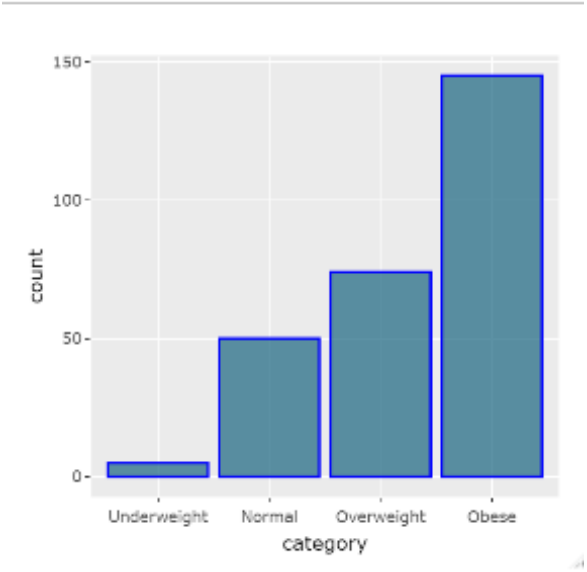
(i): Scatter Plot - Age VS charges VS bmi_group

The charges billed as a function of age for the population segregated as per bmi was evaluated using a Scatter plot and are presented below.



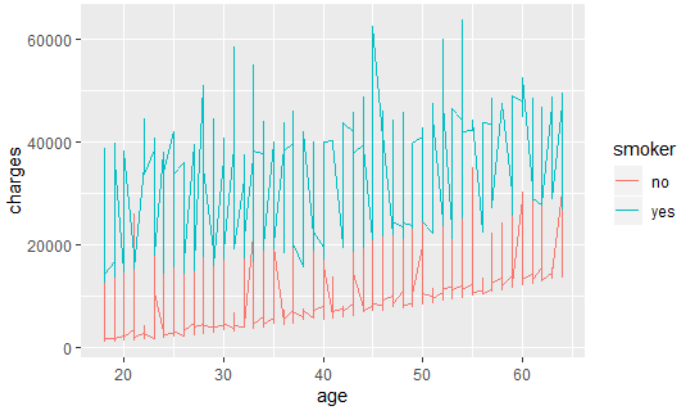
(ii):Bar Graph - Category VS Count

On the basis of bmi and smoker="yes", individuals were segregated in 4 categories namely,Underweight(bmi<18.5),Normal (bmi>18.5 and <25), Overweight(bmi>25 and <30), Obese(bmi>=30)



(iii):Line Chart - smoker VS age VS charges

The charges billed as a function of age for the smokers and non-smokers was evaluated using a line graph and are presented below.



Conclusion 1:

From the output presented above, multiple MAJOR factors like age, bmi and smoking were concluded to contribute to the charges billed for Health Insurance. Further,

(i)While the overweight individuals show a distribution of charges from $0 to $60,000, the normal population have charges consistently below $20,000.

(ii)Individuals who are smoking, are lineraly correlated with bmi which can be clearly seen by an increase in the frequency of individuals who smoke with increasing bmi.

(iii) While the charges consistently increase with increasing age for the entire population, the charges are always higher for the smokers as compared to non-smokers.

Problem Statement 2:

Construct a Joint Distribution function for age-group VS children

Output 2:

Joint Frequency Table

	0	1	2	3	4	5
>=10 <20	711	461	377	294	162	155
>=20 <30	854	604	520	437	305	298
>=30 <40	831	581	497	414	282	275
>=40 <50	853	603	519	436	304	297
>=50 <60	845	595	511	428	296	289
>=60 <70	688	438	354	271	139	132

Joint Probability Table

	0	1	2	3	4	5
>=10 <20	0.04428	0.02871	0.02348	0.01831	0.01009	0.00965
>=20 <30	0.05319	0.03762	0.03239	0.02722	0.01900	0.01856
>=30 <40	0.05176	0.03619	0.03095	0.02578	0.01756	0.01713
>=40 <50	0.05313	0.03756	0.03232	0.02715	0.01893	0.01850
>=50 <60	0.05263	0.03706	0.03183	0.02666	0.01844	0.01800
>=60 <70	0.04285	0.02728	0.02205	0.01688	0.00866	0.00822

Conclusion 2:

Joint Frequency is used to calculate how frequently the two entities are in use in the given dataset. In this scenario, the Joint Probability is calculated amongst the age-group intervals and the number of children belonging to that specific interval. Outer() function is used to create a frequency matrix with 1st 2 arguments as arrays and 3rd argument – what function is used to represent the two arrays. In order to get the probabilities from the joint frequency, we can divide the frequency matrix by the sum of its elements.

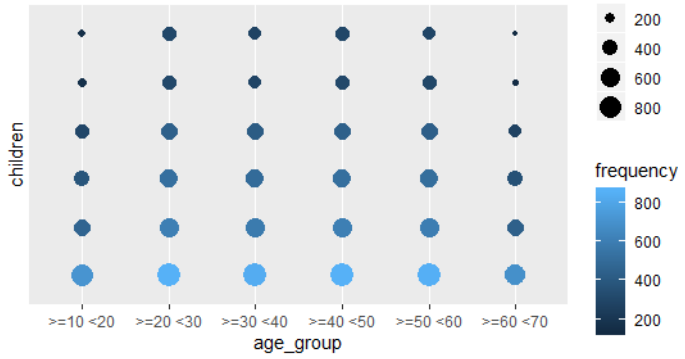
Problem Statement 3:

Perform correlation analysis on joint distribution function performed in problem 2. Determine if age and # of children are independent or dependent.

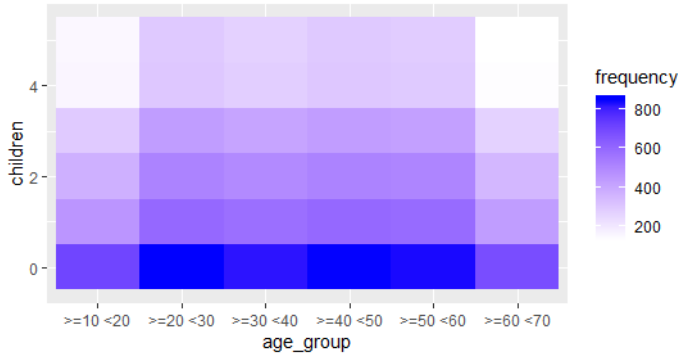
Output 3:

Corelation coefficient for age and # of children = -0.1840299

Scatter Plot



Heat Map



Conclusion 3:

On observing the heat map and the scatter plot graphs, we conclude that **age of the individuals and their # of children are independent of each other.** Further, from the scatter plot we conclude that the age intervals of 0-10 and 60-70 have almost the same distribution which doesn't make much sense.

Problem Statement 4:

Test of Hypothesis for the difference between the average charge of individuals with and without children in a particular region (Southeast)

Output:

Let X1 be the R.V. of average charge of individuals with no children

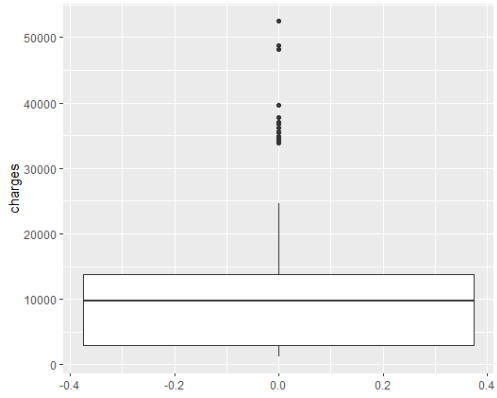
Let X2 be the R.V. of average charge of individuals with children

Our null and alternate hypothesis are:

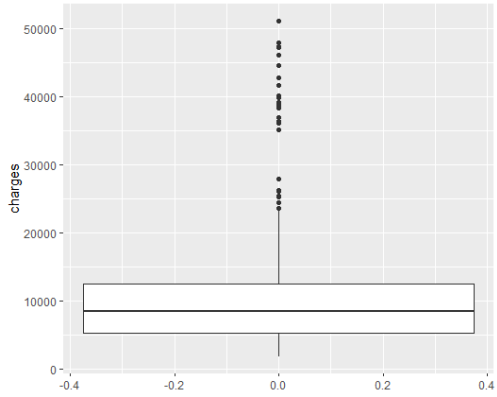
*H0 : $\mu_1 - \mu_2 = 0$

*H1 : $\mu_1 - \mu_2 \neq 0$

Z-test value = -0.5495448



Box plot Analysis for Children in Southwest with respect to charges



Box plot Analysis for Children in Southeast with respect to charges

As can be seen in the box plots above, the median lies at 10,000 and the Quartile 3 is at ~ 25,000 for both (with or with no children) the cases. Hence, we conclude that there is no significant difference between the charges for the two cases presented above.

Concluion 4:

Thus, for a significance level of α= 0.05, **we fail to reject the null hypothesis** since the z-value lies within the range [−1.96, 1.96] and conclude that there is no significant difference between the average charges of two individuals with and without children.

Problem Statement 5:

Perform Test of Hypothesis (TOH) to compare the ratio of the people who smoke in 2 different regions.

Output 5:

Let X1 be the R.V. of individuals who smoke in "southwest" region.

Let X2 be the R.V. of individuals who smoke in "southeast" region.

Our null and alternate hypothesis are:

• H0 : $p_1 = p_2$

• H1 : $p_1 \neq p_2$

data: c(91, 58) out of c(1338, 1338)

X-squared = 7.2777, df = 1, p-value = 0.006982

alternative hypothesis: two.sided

95 percent confidence interval: 0.006565644 0.042761710

sample estimates: prop 1 prop 2 0.06801196 0.04334828

P-value = 0.006981561

Conclusion 5:

The p-value of prop-test is p = 0.0069 which is smaller than the significance level 0.05. Hence, **we reject the Null Hypothesis.** Further, the proportion of smokers in southwest region is not equal to that of the southeast region.

Problem Statement 6:

Test for ratio of variances across all the regions for beneficiaries who are younger than 50 years and older than 50 years.

Output 6:

Let X1 be the R.V. of individuals who are younger than 50 years.

Let X2 be the R.V. of individuals who are older than 50 years.

Our null and alternate hypothesis are:

• H0 : $\sigma_A^2 = \sigma_B^2$

• H1 : $\sigma_A^2 \neq \sigma_B^2$

F test to compare two variances

data: *var1-age and var2-age* F = 0.16608, num df = 355, denom df = 952, **p-value < 2.2e-16**

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval: 0.1402354 0.1980472

sample estimates: ratio of variances = 0.1660797

Conclusion 6:

The p-value of F-test is ~0, which is lesser than the significance level 0.05. Thus **we reject the null hypothesis.** Further, the variances of the individuals who are younger than 50 years are not equal to the variances of individuals greater than 50 years.

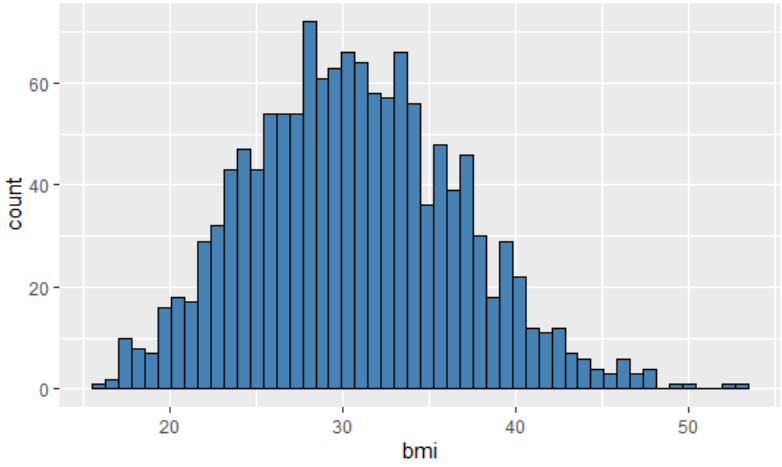
Problem Statement 7:

Estimating the distribution fitting of bmi (Continuous Data) and children (Discrete Data) along with parameter estimates using Data Visualization and functions from fitdistrplus package.

Output 7:

For Continuous Data (BMI)

1. Visualizing Data : From observation, it seems that bmi-Continuous data has a right tail with a normal like distribution which is slightly Left skewed.



Histogram for BMI

2. Descriptive Analysis: Summary Statistics

min: 15.96 max: 53.13

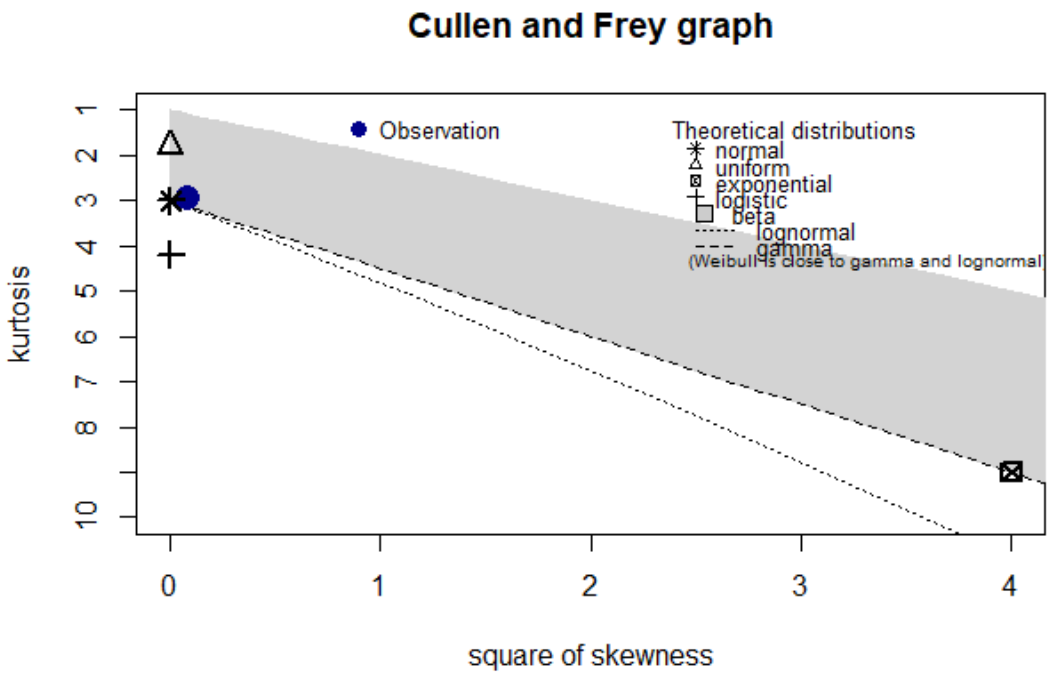
median: 30.4

mean: 30.6634

estimated sd: 6.098187

estimated skewness: 0.2840471

estimated kurtosis: 2.949268

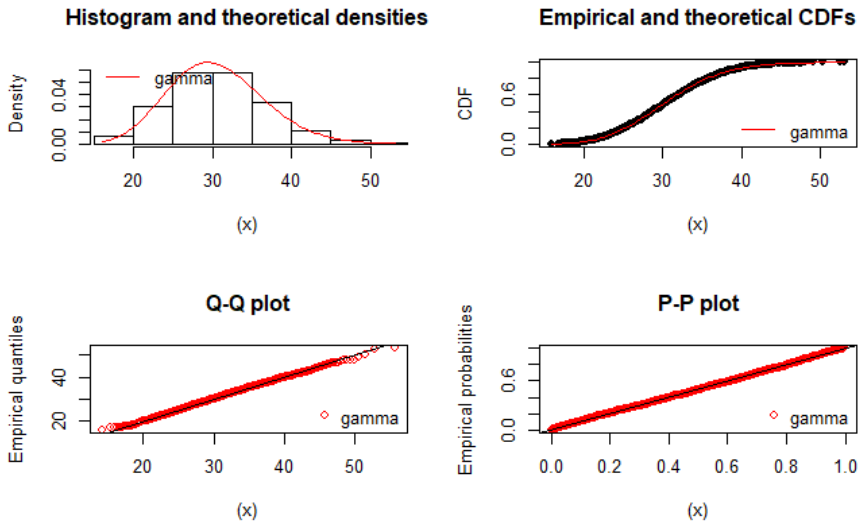


As can be observed from the plot, *gamma distribution* is the closest to the distribution for BMI and next is normal and lognormal distribution. 3.Fit: Fitting of the distribution ' gamma ' by maximum likelihood Parameters : estimate Std. Error shape 25.017624 0.96081161 rate 0.815914 0.03165116

Loglikelihood: -4306.857 AIC: 8617.713 BIC: 8628.111

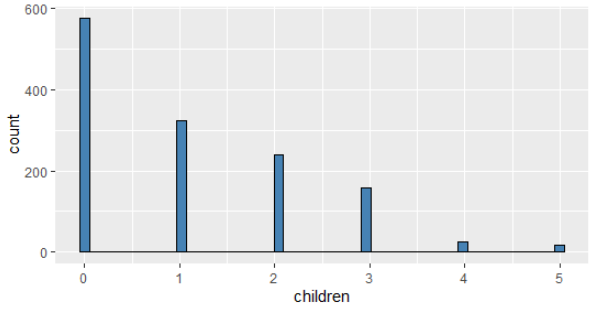
Correlation matrix: shape rate shape 1.0000000 0.9900242 rate 0.9900242 1.0000000

For Gamma Distribution: Goodness of fit



For Discrete Data (No of Children)

1. Visualizing Data : From observation, it seems that No of Children (discrete data) is normally distributed but slightly skewed to the left (right tail)



2. Fit : Negative binomial Distribution Fitting of the distribution ' nbinom ' by maximum likelihood Parameters :

   estimate Std. Error

size 2.553715 0.40094128 mu 1.094833 0.03418986

Loglikelihood: -1910.864 AIC: 3825.728 BIC: 3836.126

Correlation matrix: size mu size 1.0000000000 0.0001169998 mu 0.0001169998 1.0000000000

3. Goodness of fit: Chi-squared statistic: 64.57175 133.0106

Degree of freedom of the Chi-squared distribution: 2 3

Chi-squared p-value: 9.515302e-15 1.21405e-28

Goodness-of-fit criteria 1-mle-nbinom 2-mle-pois

Akaike's Information Criterion 3825.728 3892.885

Bayesian Information Criterion 3836.126 3898.084

Conclusion 7:

1.For Continuous Data : Skewness and Kurtosis are especially important here, as a non-zero skewness reveals a lack of symmetry, while the kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3. From these, and since AIC and BIC values are smaller and Loglikelihood is higher for Gamma distribtuion, it is evident that **Gamma distribution is the best fit** for the bmi feature with a $\mu = 30.66$ and $\sigma = 6.09$.

2.For Discrete Data : We see that **negative binomial distribution fits the data well** as the chi square statistics values are smaller in 1st case. Hence nbinom is used to fit the data well (This does not mean that the data follows these distributions with the parameters obtained using the fitdist() function, but that it cannot be distinguished from the above distributions with the parameters obtained).

Problem Statement 8:

Conduct a test a hypothesis for ratio of variances across all the regions for those who have normal bmi and those who does not have normal bmi .

Output 8:

Let X1 be the R.V. of individuals who have normal bmi(bmi >18.5 and bmi<24.9).

Let X2 be the R.V. of individuals who are either underweight or overweight (do not have normal bmi).

Our null and alternate hypothesis are:

- H0 : $\sigma_A^2 = \sigma_B^2$
- H1 : $\sigma_A^2 \neq \sigma_B^2$

F test to compare two variances

data: a-age and b-age

F = 0.9399, num df = 221, denom df = 1115, **p-value = 0.5703**

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval: 0.7715076 1.1613906

sample estimates: ratio of variances = 0.9399034

Conclusion 8:

The p-value of F-test is 0.5 which is greater than the significance level 0.05. As there is no significant difference between the individuals with normal bmi and those who do not fall under normal bmi criteria **we fail to reject the null hypothesis test**.