Created by Rahman Aziz Firmansyah
AI & Machine Learning Engineering Student at PACMANN

Short Report

Intro to Probability and Statistics: Final Project

In this project, a number of features given in the insurance data set were analyzed through descriptive statistics, categorical variable analysis, numerical variable analysis, correlation analysis, and hypothesis testing and several statements were claimed true supported by statistical evidences as follows.

- Among numerical features in the data set, users' **age** is the **most correlated** feature with the amount of **charges** paid by users
- **Smoker** users pay **higher charges** than that of non smoker users
- Users with **BMI over 25** pay **higher** than that of users with BMI less than 25
- Average BMI of male users is **higher** than that of female

Descriptive Statistics

In this section, a number of analyses were performed to answer the questions and identify the characteristics and feature distribution.

Q1: What is the average users' age?
A1: Without considering the distribution, we can give a quick answer that the average users' age was **39 years old.**

Q2: What is the average BMI of users who smoke or do not smoke?
A2: By using the average formula, the calculated average users' BMI are **30.71 (given a smoker)** and **30.65 (given a nonsmoker)**. Average BMI value of a smoker is slightly higher than that of a nonsmoker.

Q3: What are the average charges of smoker users and nonsmokers users?
A3: The average charges of **smoker** users is **32050.23** and the average charges of **nonsmoker** users is **8434.27.** The average charges applied to smoker users is significantly higher than that of nonsmoker.

Q4: What is the comparison between average BMI given a user aged over 25 years old and smoker and user aged over 25 and nonsmoker?
A4:

$$\bar{x}_{(BMI|age>25 \cap smoker=1)} = 30.58$$

$$\bar{x}_{(BMI|age>25 \cap smoker=0)} = 30.93$$

Average BMI of users who is aged over 25 and nonsmoker is **slightly higher** than that of smoker.

Q5: What is the comparison between average BMI of male and female users?
A5:

$$\bar{x}_{(BMI|male)} = 30.94$$

$$\bar{x}_{(BMI|female)} = 30.38$$

Average BMI of male users is **slightly higher** than that of female.

Categorical Variable Analysis

Q1: What are the proportions of smoker and nonsmoker in the data set?
A1:

% Users who smoke= 20.48%
% Users who do not smoke = 79.52%

There are **more users who do not smoke** than users who smoke.

Q2: Does each region has identical proportion of users?
A2:

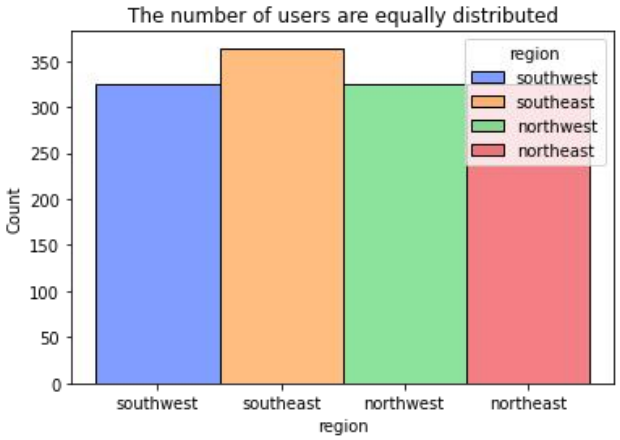

Image 2.1. The distribution of insurance users In each region

Each region **has identical proportion** of users even though a slight difference is observed in south east region.

Q3: Which category has a higher proportion? Smoker or nonsmoker?
A3: The proportion of nonsmoker is higher than smoker.

Q4: What is the probability of a user is male given a smoker?
A4: The probability of a user is male given a smoker is 0.58

Q5: What is the probability of a user is female given a smoker?
A5: The probability of a user is female given a smoker is 0.42

Created by Rahman Aziz Firmansyah
AI & Machine Learning Engineering Student at PACMANN

Continuous Variable Analysis

Q1: How does the probability of charges applied to users look like

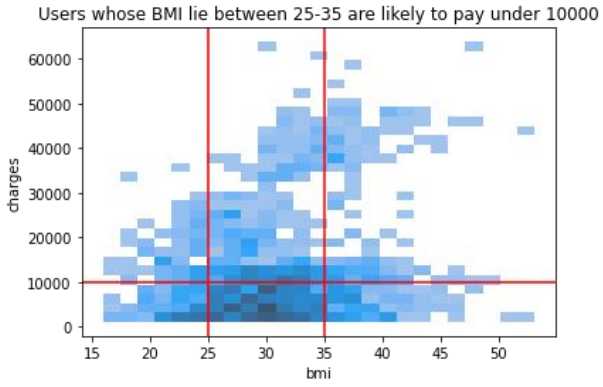

Image 3.1. The distribution of charges with respect to BMI

Q2: What is the probability of a smoker whose BMI is over 25 will be charged more than 16.700?
A2:
Let be X = charges, Y = BMI, Z = smoker (takes two states, yes (1) or no (0))

$$P(X > 16700 | Y > 25 \cap Z = 1) = \frac{P(X > 16700 \cap (Y > 25 \cap Z = 1))}{P(Y > 25 \cap Z = 1)} = 0.98$$

Q3: What is the probability of a smoker will be charged more than 16.700?
A3:

$$P(X > 16700 | Z = 1) = \frac{P(X > 16700 \cap Z = 1)}{P(Z = 1)} = 0.93$$

Q4: Which event is more likely to happen, a user whose BMI over 25 being charged more than 16.7k or a user whose BMI under 25 being charged more than 16.7k?
A4:

$$P(X > 16700 | Y > 25) = \frac{P(X > 16700 \cap Y > 25)}{P(Y > 25)} = 0.26$$

$$P(X > 16700 | Y < 25) = \frac{P(X > 16700 \cap Y < 25)}{P(Y < 25)} = 0.21$$

A user whose BMI greater than 25 is more likely to be charged over 16.7k.

Q5: Which event is more likely to happen, a smoker with BMI over 25 being charged more than 16.7k or a nonsmoker with BMI under 25 being charged more than 16.7k?
A5:

$$P(X > 16700 | Y > 25 \cap Z = 0) = \frac{P(X > 16700 \cap (Y > 25 \cap Z = 0))}{P(Y > 25 \cap Z = 0)} = 0.08$$

$$P(X > 16700 | Y > 25 \cap Z = 1) = \frac{P(X > 16700 \cap (Y > 25 \cap Z = 1))}{P(Y > 25 \cap Z = 1)} = 0.08$$

Interestingly, both events are equally unlikely to happen.

Correlation Analysis

In this section correlation between continuous features were calculated and mapped. Users' age is most correlated among continuous variables despite there is no strong correlation observed.

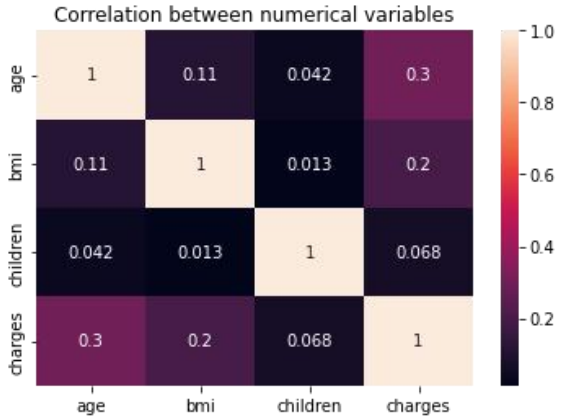

Image 4.1 Correlation between continuous variables

Correlation strength to charges in descending order: age > BMI > number of children

Hypothesis Testing

In this section, hypothesis testing was performed to validate statements obtained in previous sections and provide statistical evidence to the claim.

Q1: Are the charges applied to smokers higher than that of nonsmokers?
A1:
Independent t-test (left-tailed) was performed to provide statistical evidence upon our claim

$$H_0 : \bar{x}_{smoker} \geq \bar{x}_{nonsmoker}$$

$$H_1 : \bar{x}_{smoker} < \bar{x}_{nonsmoker}$$

Since p-value obtained from t-test is greater than alpha, we **failed to reject the null hypothesis** and we can conclude that **charges applied to smokers is higher than that of nonsmokers.**

Q2: Are the charges for users with BMI over 25 is higher than that of users with BMI less than 25?
A2:

$$H_0 : \bar{x}_{BMI>25} \leq \bar{x}_{BMI\leq 25}$$

$$H_1 : \bar{x}_{BMI>25} > \bar{x}_{BMI\leq 25}$$

P-value obtained is less than alpha, **the null hypothesis is rejected** and we should claim that average charges applied to users with BMI over 25 is higher.

Q3: Are BMI of male and female identical?

A3:

$$H_0 : \bar{x}_{male} \leq \bar{x}_{female}$$

$$H_1 : \bar{x}_{male} > \bar{x}_{female}$$

P-value calculated is less than alpha, thus the null hypothesis is rejected and we should claim that average male BMI is higher than that of female.