**Task 2: Data Exploration**

Joseph Emmanuel Cayetano

Western Governors University

D599: Data Preparation and Exploration Task 2

October 14, 2024

1. **Identify the distribution of two continuous variables and two categorical variables using univariate statistics from the dataset**

The two continuous variables that I have chosen in the dataset are ‘charges’ and ‘bmi’. For these variables, I used a univariate statistic method which is the .describe() function to provide a descriptive summary for each continuous variables including count, mean, standard deviation, minimum, 25%, 50%, 75%, and maximum.

A screenshot of a computer

Description automatically generated

These are the summary statistics for ‘charges’. It has a count of 1338, mean of 13270.42, standard deviation of 12110.01, minimum of 1121.87, 25% of 4740.28, 50% of 9382.03, 75% of 16639.91, and a maximum of 63770.42. The ‘charges’ variable has a float data type, and it does not contain any missing values.

A screenshot of a computer

Description automatically generated

These are the summary statistics for ‘bmi’. It has a count of 1338, mean of 30.66, standard deviation of 6.09, minimum of 15.96, 25% of 26.29, 50% of 30.4, 75% of 34.69, and a maximum of 53.13. The ‘bmi’ variable has a float data type, and it does not contain any missing values.

The two categorical variables that I have chosen in the dataset are ‘sex’ and ‘smoker’. For these variables, I used a univariate statistic method which is the .value\_counts() function to provide frequency analysis for each categorical variable. The .value\_counts() function will output the total count of each unique value in a categorical variable. I converted the counts into proportions.

A screenshot of a computer code

Description automatically generated

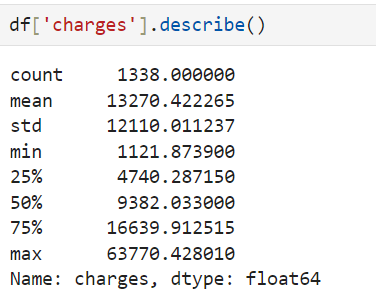
This is the proportion of each unique value in the ‘sex’ variable. I used the argument ‘normalize=True’ to return the relative frequency of each value instead of total counts. The proportion of males is 50.52%, while the proportion of females is 49.48%. There are more males than females in the dataset but by a very slim margin. The ‘sex’ variable also has a float data type.

A screenshot of a computer code

Description automatically generated

This is the proportion of each unique value in the ‘smoker’ variable. I used the argument ‘normalize=True’ to return the relative frequency of each value instead of total counts. The proportion of non-smokers is 79.52%, while the proportion of smokers is 20.47%. There are significantly more non-smokers than smokers in the dataset. The ‘smoker’ variable also has a float data type.

**A1. Represent your findings from part A visually as part of your submission**



A screenshot of a graph

Description automatically generated

For the continuous variable ‘charges’, I used a boxplot to visually represent the distribution and summary statistics of the variable shown in the describe() output. The descriptive statistics of the ‘charges’ variable can be seen on the boxplot. The boxplot is right-skewed, and it has a long tail of higher charges and many outliers.

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

For the continuous variable ‘bmi’, I used a boxplot to visually represent the distribution and summary statistics of the variable shown in the describe() output. The descriptive statistics of the ‘bmi’ variable can be seen on the boxplot. The boxplot is symmetrical, but it has a few high outliers that suggests some people have significantly higher BMIs compared to the average.

A screenshot of a computer code

Description automatically generated

A screen shot of a computer

Description automatically generated

For the categorical variable ‘sex’, I used a pie chart to visually represent the relative frequency of each unique value in the variable as shown in the value\_counts() output. The proportions of males and females can be seen on the pie chart. Males slightly outnumber females in the data.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

For the categorical variable ‘smoker, I used a pie chart to visually represent the relative frequency of each unique value in the variable as shown in the value\_counts() output. The proportions of smokers and non-smokers can be seen on the pie chart. Non-smokers outnumber smokers in the data by a large margin.

1. **Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from the dataset**

A screenshot of a phone

Description automatically generated

The two continuous variables that I have chosen are ‘bmi’ and ‘charges’. The bivariate statistics method that I used is the Pearson correlation coefficient. The reason for using the Pearson correlation coefficient is to see if there is a correlation between ‘bmi’ and ‘charges’. The correlation coefficient between the two variables is 0.19 or 19%. This indicates a positive but weak linear relationship between the two continuous variables.

A screenshot of a computer

Description automatically generated

The two categorical variables that I have chosen are ‘sex’ and ‘smoker’. The bivariate statistics method that I used is cross-tabulation. The reason for using cross-tabulation is to see the proportion of smokers and non-smokers within each sex category. For females, 82% are non-smokers while 18% are smokers. For males, 76% are non-smokers, while 23% are smokers. Overall, a higher percentage of females are non-smokers compared to males, and a higher percentage of males are smokers compared to females.

A screenshot of a computer

Description automatically generated

I have also performed a bivariate statistics method between a categorical variable and a continuous variable. The bivariate statistics method that I used is the groupby() function. The reason for using the groupby() function is to perform a descriptive analysis to compare insurance charges based on smoking status. The average insurance charge for a smoker is 32050.23, while the average insurance charge for a non-smoker is 8434.26. Overall, smokers have higher insurance charges than non-smokers on average.

**B1. Represent your findings from part B visually as part of your submission**

A screenshot of a phone

Description automatically generated

A screen shot of a screen

Description automatically generated

I used a scatterplot to represent the correlation coefficient between two continuous variables, which are ‘bmi’ and ‘charges’. The scatterplot shows a positive but weak relationship between ‘bmi’ and ‘charges’. This indicates that BMI is not a strong predictor of insurance charges in the data.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

I used a bar chart to represent the cross-tabulation chart for the two categorical variables, which are ‘sex’ and ‘smoker’. The bar chart shows the proportion of smokers and non-smokers by sex. There are more male smokers than female smokers, and more female non-smokers than male non-smokers. Essentially, smoking is more common males than females.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

I used a box plot to represent the descriptive analysis that compare insurance charges based on smoking status. The box plot shows that smokers pay significantly higher insurance chargers compared to non-smokers, on average. This indicates that smoking is linked to higher medical costs, likely driven by increased health risks.

**C1. Provide one research question relevant to the dataset and any organizational needs that can be answered through data analysis**

Research Question: Is there a significant difference in the mean BMI between males and females?

**C2. Identify the variables in the dataset that are relevant to answering the research question**

To answer the research question, the variables in the dataset that need to be used are 'bmi' and 'sex'. The 'bmi' variable is a continuous variable and it measures the body mass index. The 'sex' variable is a categorical variable that has two groups: males and females. The mean BMI of males will be compared to the mean BMI of females to see if there is a significant difference in BMI between the two groups.

**D1. Identify a parametric statistical test that is relevant to your question**

The parametric statistical test that will be used to answer the research question is the independent t-test. The independent t-test is “a statistical method of hypothesis testing that determines whether there is a statistically significant difference between the means of two independent samples” (Patel, 2020, par. 1). This test perfectly fits with our research question because we want to know if there is a significant difference in the average BMIs between the two groups: males and females.

**D2. Develop null and alternative hypotheses related to your chosen parametric test**

Null Hypothesis (H0): μmale = μfemale

There is no significant difference in the mean BMI between males and females.

Alternative Hypothesis (H1): μmale =/= μfemale

There is a significant difference in the mean BMI between males and females.

**D3. Write code (in either Python or R) to run the parametric test**

A screenshot of a computer code

Description automatically generated

In the code, the BMI values for females is being stored into the variable 'bmi\_females', while the BMI for males is being stored into the variable 'bmi\_males'. We're performing the independent t-test by using the ttest\_ind() function to compare the BMI means of males and females. The ttest\_ind() function will return two values, which are t-statistic and p\_value.

**D4. Provide the output and the results of any calculations from the parametric statistical test you performed**

A number with black text

Description automatically generated with medium confidence

Since the p-value is more than the 0.05 significance level, we fail to reject the null hypothesis and conclude that there is no significant difference in the mean BMI between males and females. Regarding the t-statistic value, it indicates that the difference in BMI means between males and females is 1.6967 standard errors away from zero. Since the value of t-statistic is negative, it indicates that females have slightly higher BMI than males.

**E1. Justify why you chose the statistical test identified in part D1 based on variables**

I chose to use independent t-test because it compares the means of a continuous variable between two unrelated groups. Since the variable 'bmi' is a continuous variable and the variable 'sex' is a categorical variable with two groups: males and females, it is suitable to use independent t-test with those variables.

**E2.** **Discuss the test results, including the decision to reject or fail to reject the null hypothesis**

After performing the independent t-test involving the BMIs of males and females, it returns a T-statistic value of -1.69 and a p-value of 0.08. The T-statistic value means there is a moderate difference between the BMI means of males and females, and since it is negative, it indicates that females have slightly higher BMI values than males. The p-value is 0.08 which is less than the 0.05 significance level. This means we fail to reject the null hypothesis and conclude there is no significant difference in BMI between males and females.

**E3. Explain how stakeholders in the organization benefit from your choice of testing method**

Using the independent t-test on variables 'bmi' and 'sex', it benefits the stakeholders in a couple of ways. First, if a significant difference is found in BMI between males and females, stakeholders can use this result to design targeted health programs for the group with higher BMI, leading to better health outcomes. Second, using a parametric test like independent t-test allows organizations to develop health strategies with the help of actual data instead of assumptions.

**F1. Discuss the answer to your question from part C1**

Based on the result from the independent t-test, the answer to the question "Is there a significant difference in the mean BMI between males and females?" is no. There is no significant difference in the mean BMI between males and females at the 0.05 significance level. The p-value (0.089) is greater than 0.05, so we failed to reject the null hypothesis. It is safe to conclude that sex alone may not be a key factor in influencing BMI at least in this dataset.

**F2. Discuss the limitations of your data analysis**

There are limitations to my data analysis. The first limitation is I did not get to check the BMI variable for outliers. Independent t-tests are sensitive to outliers, so by not removing the outliers from the 'bmi' column, the outliers can affect the mean and potentially skew the results. Another limitation is the independent t-test only analyzes the relationship between two variables without considering other variables that might influence the results. In our case, only the variables 'sex' and 'bmi' were analyzed by the independent t-test and nothing else. It did not consider any variables like age or region in the data where it could influence BMI.

**F3. Recommend a course of action based on your findings**

Based on the findings, we can follow some recommended courses of actions. First, since there is no significant difference in BMI between males and females, there is no need to create gender-specific health and wellness programs based solely on BMI. Both genders can be treated similarly regarding BMI in health plans and programs. Second, it would be better for organizations to focus on developing programs that emphasize healthy habits for everyone, regardless of gender.

**F4. Submit your code for each test**

The code that I used to conduct parametric statistical testing is in the "D599\_Task\_2.ipynb" file.

**G1. Provide one research question relevant to the dataset and any organizational needs that can be answered through data analysis**

Research Question: Do smokers have significantly higher insurance charges than non-smokers?

**G2. Identify the variables in the dataset that are relevant to answering your research question**

To answer the research question, the variables in the dataset that need to be used are 'smoker' and 'charges'. The 'charges' variable is a continuous variable, and it contains the insurance charges for each individual. The 'smoker' variable is a categorical variable that has two groups: smokers and non-smokers. The distribution of insurance charges for smokers will be compared to the distribution of insurance charges for non-smokers to see if smokers have significantly higher insurance charges.

**H1. Identify a nonparametric statistical test that is relevant to your question**

The non-parametric statistical test that will be used to answer the research question is the Mann-Whitney U test. The Mann-Whitney U test compares two independent groups to see if one group tends to have higher or lower values than the other group. In our case, we are using the Mann-Whitney U test to see whether smokers tend to have significantly higher insurance charges than non-smokers. The Mann-Whitney U test is also a good fit because the distributions for both smokers and non-smokers are skewed and do not follow a normal distribution.

**H2. Develop null and alternative hypotheses related to your chosen nonparametric test**

Null Hypothesis (H0): MEDsmokers = MEDnonsmokers

The distribution of insurance charges for smokers is equal to the distribution of insurance charges for non-smokers.

Alternative Hypothesis (H1): MEDsmokers > MEDnonsmokers

The distribution of insurance charges for smokers is greater than the distribution of insurance charges for non-smokers.

**H3. Write code (in either Python or R) to run the nonparametric test**

A screenshot of a computer code

Description automatically generated

In this code, the insurance charges for smokers are stored in the variable 'smokers\_charges', while the insurance charges for non-smokers are stored in the variable 'non\_smokers\_charges'. We're performing the Mann-Whitney U test by using the mannwhitneyu() function to see whether smokers tend to have significantly higher insurance charges than non-smokers. For the alternative parameter, I used 'greater' because our research question has a one-sided hypothesis where smokers are expected to have higher insurance charges than non-smokers. The mannwhitneyu() function will return two values, which are u-statistic and p\_value.

**H4. Provide the output and the results of any calculations from the nonparametric statistical test you performed**

A screenshot of a computer screen

Description automatically generated

Based on the Mann-Whitney U test results, the p value is less than the 0.05 significance level, which means we can confidently reject the null hypothesis and conclude that the distribution of insurance charges for smokers is greater than the distribution of insurance charges for non-smokers. Regarding the u-statistic value, it is quite large, which indicates that the smokers group tends to have higher insurance charges than the non-smokers group.

**I1. Justify why you chose the statistical test identified in part G1 based on variables**

I chose to use the Mann-Whitney U test for two reasons. The first reason is because the distributions for both smokers and non-smokers are skewed and do not follow a normal distribution. A non-parametric test like the Mann-Whitney U test is appropriate for skewed data like this. The second reason is about the variables. The 'smoker' variable has two independent groups: smokers and non-smokers. The Mann-Whitney U test is designed to compare two independent groups. Also, the 'charges' variable is a continuous variable, which the function uses to compare between two groups.

**I2. Discuss test results, including the decision to reject or fail to reject the null hypothesis**

After performing the Mann-Whitney U test, it returns a u-statistic value of 284133.0 and p-value of 2.6-130. Since the u-statistic value is large, it means that the values in the smokers group tend to rank higher than the values in the non-smokers group. The p-value has a value of 2.6-130 which is smaller than the 0.05 significance level. Therefore, we can reject the null hypothesis and conclude that smokers have significantly higher insurance charges than non-smokers.

**I3. Explain how stakeholders in the organization benefit from your choice of testing method**

Using the Mann-Whitney U test on variables 'smoker' and 'charges', it benefits the stakeholders in a couple of ways. First, by having this information about smokers having higher insurance charges than non-smokers, it justifies their action of increasing premium rates for high-risk groups like smokers. Second, having the test results from the Mann-Whitney U test will allow marketing to develop targeted insurance products that address the needs of smokers.

**J1. Discuss the answer to your question from part G1**

Based on the results from the Mann-Whitney U test, the answer to the question "Do smokers have significantly higher insurance charges than non-smokers?" is yes. Smokers tend to have significantly higher insurance charges than non-smokers. The extremely low p-value (2.6-130) from the test allows us to reject the null hypothesis and conclude that smoking status is associated with higher insurance charges.

**J2. Discuss the limitations of your data analysis**

There are limitations to my data analysis. The first limitation is the Mann-Whitney U test compares only two independent groups in a categorical variable. It is not possible to use the Mann-Whitney U test if you want to compare more than two independent groups. The second limitation is that the Mann-Whitney U test does not provide the exact measure of how high the insurance charges for smokers are compared to non-smokers. We have to use descriptive statistical methods like the interquartile range to really understand the magnitude of differences between the two groups.

**J3. Recommend a course of action based on your findings**

Based on the findings, we can follow some recommended courses of actions. First, organizations can implement higher premiums for smokers to reflect the increased costs associated with smokers. This helps organizations manage their costs more effectively. Second, organizations can offer discounted premiums to smokers who successfully quit smoking. This helps reduce insurance claims while also promoting healthier behaviors.

**J4. Submit your code for each test**

The code that I used to conduct non-parametric statistical testing is in the "D599\_Task\_2.ipynb" file.

**K. Panopto Video Presentation**

Here is the link to my Panopto presentation: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=686b8490-bd29-4f9a-89b3-b20a01099d0b

**References**

Patel, K. (2020, October 7). *The independent samples T-Test Method and how IT benefits organizations*. DATAVERSITY. https://www.dataversity.net/the-independent-samples-t-test-method-and-how-it-benefits-organizations/