Chi-Square Analysis – Performance Assessment

Steven Schindler

Exploratory Data Analysis – D207

Sewell, Williams; PhD

College of I.T., Western Governors University

Table of Contents

A1.

The data set I have chosen to work with is the same Churn data set that I worked with in D206. The question that I propose is “Is there a relationship between Contract and Churn?” This question will require a chi-square analysis and will have two hypotheses. The Null hypothesis which is “There is no relationship between Contract and Churn.” Then there is the research hypothesis “There is a relationship between Contract and Churn.”

A2.

The organization’s stakeholders could benefit from this analysis is that they will now know if there is a relationship between Contract and Churn. If there is one, they can dive deeper into what type of relationship is there? What factors of contract contribute to high churn? What factors contribute to low churn? A couple of examples being “Which contract has the highest churn?”, in contrast “Which contract has the lowest churn?” All these questions can be asked and answered for stakeholders if a relationship is first established between Contract and Churn.

A3.

The two variables that I need from my data set are Churn and Contract. Churn is a qualitative nominal variable that tells whether a customer has discontinued service in the past month. E.g., yes, no. The Contract variable is also a qualitative nominal variable that tells the length of the contract the customer has with the company. E.g., Month-to-month, One year, Two year.

B1.

I chose to do a chi-square analysis in the R programming language. The code to perform the chi-square analysis is as follows:

```
getwd()
library(dplyr)

set.seed(123)
df <- read.csv('churn_clean.csv', header = TRUE)

churn<-df$Churn
contract<-df$Contract
chisq.test(contract, churn)
```

First, I get a look at the current working directory. Then I load the dplyr library and set the seed to 123. I then read the csv file into a dataframe called “df” I then assign the columns of Churn and Contract to variables churn and contract respectively. Then I use the built in chi-square function to perform chi-square analysis on contract and churn.

B2.

The output from the code is as follows:

```
Pearson's Chi-squared test

data:  contract and churn
X-squared = 718.59, df = 2, p-value < 2.2e-16
```

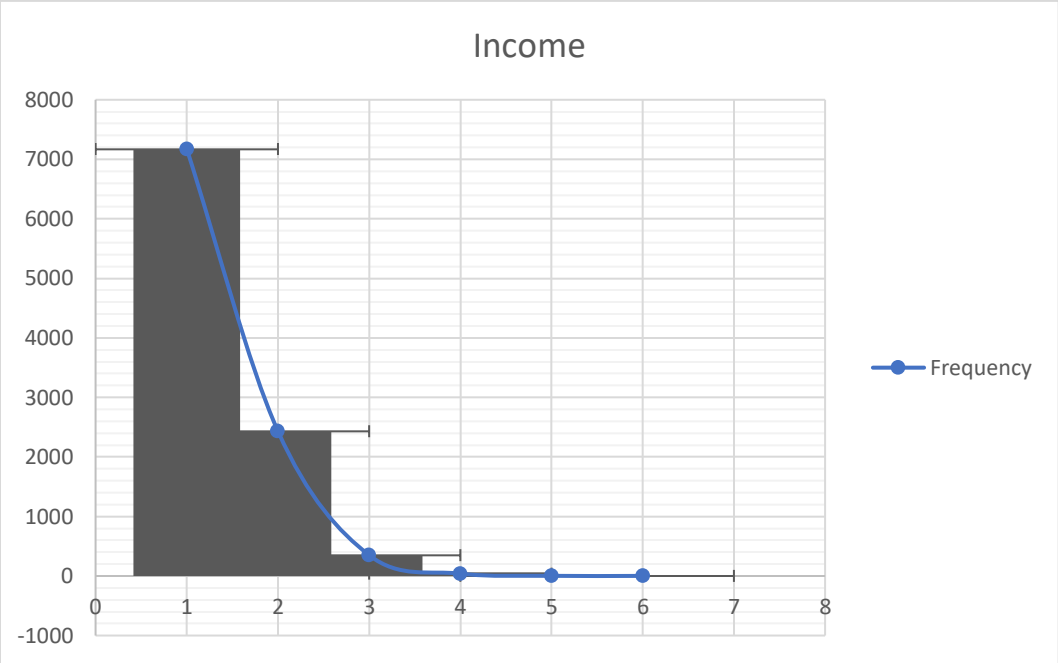
The df on the bottom row is degrees of freedom and is 2 while the p-value or alpha value is 2.2 * 10^-16 much less that .05. The chi-square value is 718.59.

B3.

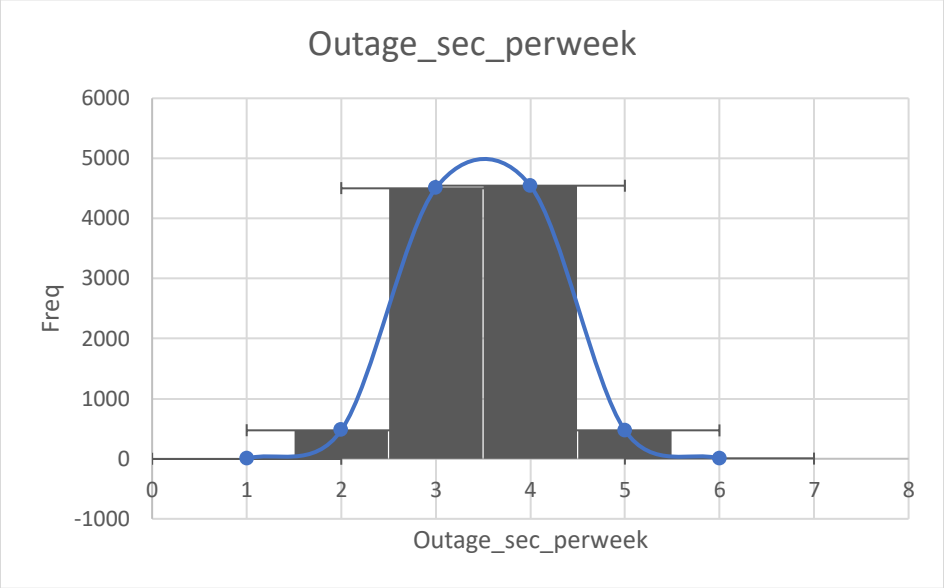
I chose to use a chi-square analysis because both variables that I used(Contract and Churn) were categorical and the chi-square method is best for finding relationships between two categorical variables. This will help me answer the research question in A1 of finding a relationship between Contract and Churn.

C1.

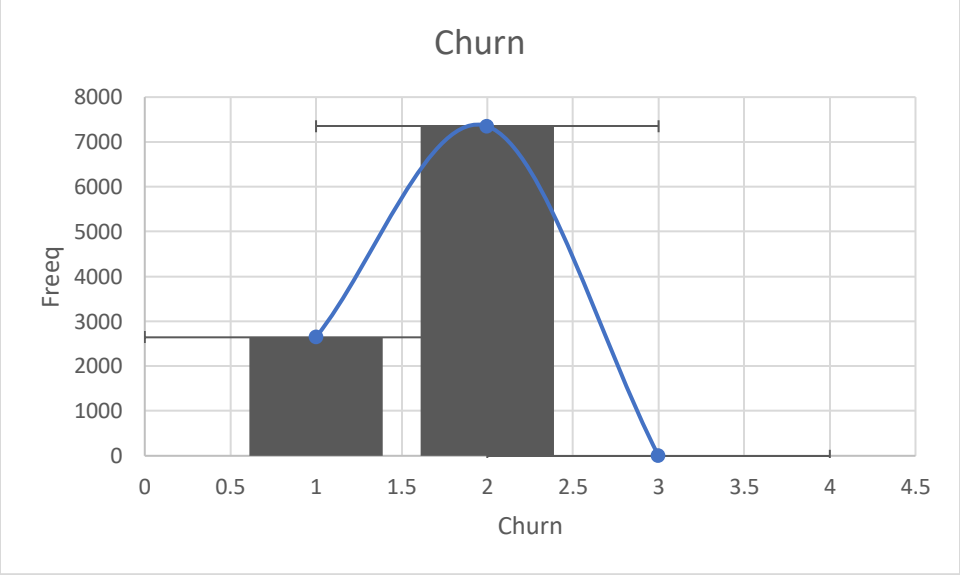
The first distribution is of Income (continuous), and it is a right skewed lognormal distribution. Below is a histogram of Income title “Income” that shows the distribution is lognormal and right skewed. This graph and all the rest of the graphs in C1 were made using Excel.



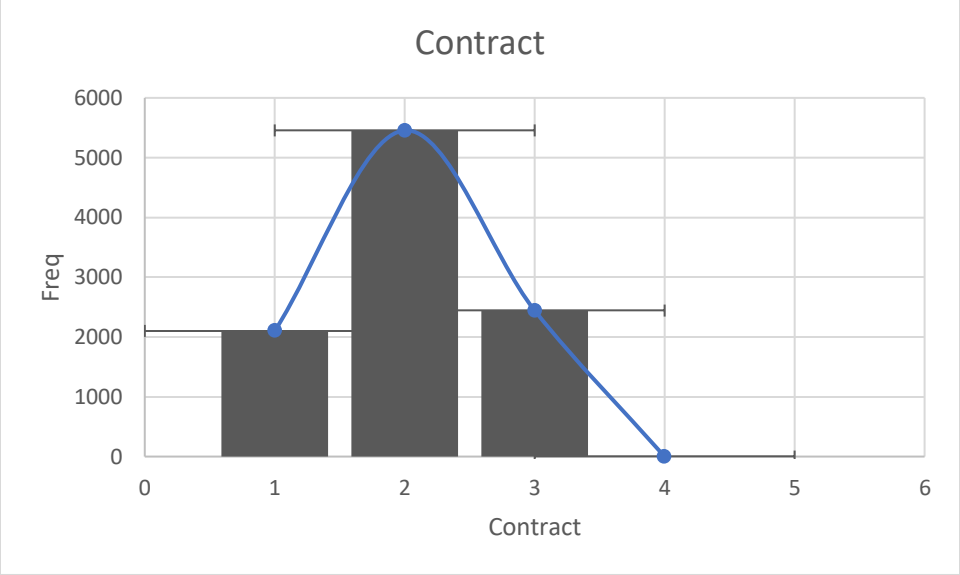
The next continuous variable is Outage_sec_perweek which is a normal distribution with a bell curve shape. It is pictured below in a histogram titled “Outage_sec_perweek”



The first categorical variable is Churn and it is a discrete Poisson distribution (Investopedia) as it is looking at the amount of churn per month. It is below in the histogram titled “Churn”. Since it is categorical, I converted the values to numerical with a 1 being a “yes” and a 2 being a “no”.

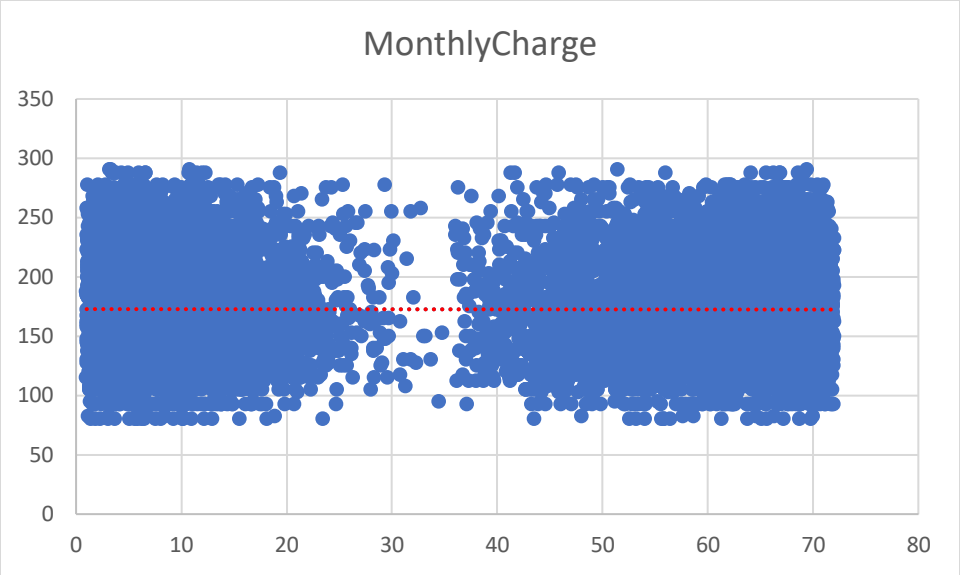


The final categorical variable is a discrete multinomial distribution. (Investopedia). It is shown below with the histogram title “Contract”. Like with the Churn variable Contract will also need to be converted to numeric so 1 is “one year”, 2 is “month-to-month” and 3 is “two year”.

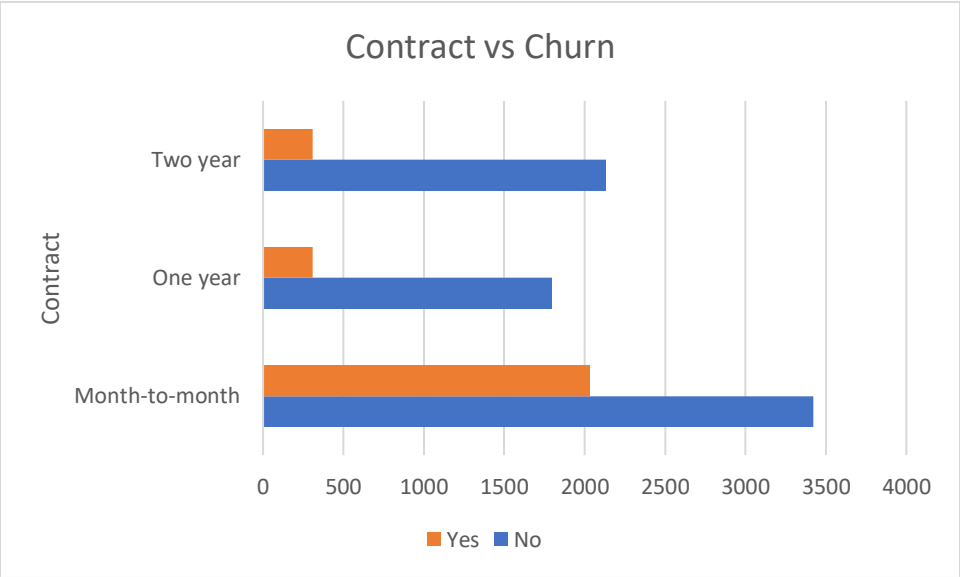


D1.

The two continuous bivariate variables I chose were Tenure and MonthlyCharge. They are represented in the scatterplot below labeled “Tenure vs. MonthlyCharge”. The red linear shows a horizontal line indicating a slope of zero and therefore little correlation between Tenure and MonthlyCharge. This graph and the other graph in D1 were made using Excel.



The two categorical variables I chose were Contract and Churn. I represented them in a clustered bar chart below. From the chart we can see that Month-to-month contracts have the highest amount of churn with around 2000 yeses. While one-year and two-year "yes" counts are very close together.



E1.

From the output given in B2 we know that our or p-value of 2.2 * 10^-16 is less than an alpha level of .05 thus we can infer there is a statistically significant relationship between Churn and Contract. We can also see from the chi-square distribution table(Nazarathy, Y., 2007) that with 2 degrees of freedom and an alpha level of .05 we need X^2 value of at least 5.99. The X^2 value given by the R output was 718.59 which is greater than 5.99 giving further evidence to the strong relationship between Churn and Contract. I therefore can reject the Null hypothesis and conclude there is a relationship between Churn and Contract according to this preliminary analysis.

E2.

The main limitation of the chi-square analysis is that it cannot find casual relationships between variables it can only determine if two variables are independent of each other. Another limitation is that it is sensitive to sample size with a big enough sample size any variables may have a relationship that looks statistically significant when, in reality they are independent of each other. (University of Utah, Department of Sociology.)

E3.

The original research question was “Is there a relationship between Contract and Churn?” since I could reject the Null hypothesis the organization can now investigate the significance of a relationship between Contract and Churn . They can now investigate such questions as What factors of contract contribute to high churn? What factors contribute to low churn? A couple of examples being “Which contract has the highest churn?”, in contrast “Which contract has the lowest

churn?" In short, I recommend that the company does more analysis on ways to reduce Churn involving Contract.

G.

No other sources were used for third- party code.

H.

Nazarathy, Y. (2007). The chi-squared distribution table. Retrieved March 30, 2023, from https://people.smp.uq.edu.au/YoniNazarathy/stat_models_B_course_spring_07/distributions/chisqtab.pdf

University of Utah, Department of Sociology. (n.d.). Chi-square. Retrieved March 30, 2023, from https://soc.utah.edu/sociology3112/chi-square.php