**Chi-square Model on Shoe Categories**

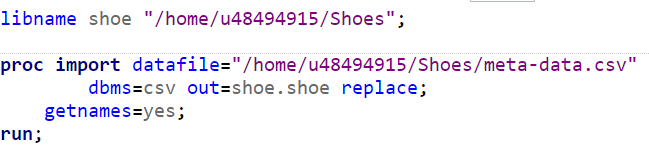
**Executive Summary**

The online shoe retailer Zappos is identified as having the best customer service in the world. This study aims to create a chi-square test to compare the significant differences among the shoe categories from the online catalog of Zappos.com. The analysis aims to compare the relationship between leather shoes and sales. Since leather is an expensive material, if shoes do not show a significant capacity to sell, those shoes could be manufactured with a different material less expensive than leather. The analysis will be to compare the shoe categories against the shoe material. The hypothesis derived from the project's goal is summarized as follows: There is no significant difference between the material and the shoe category.

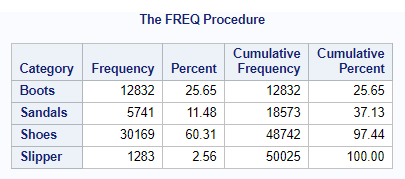
UT-Zap50K is a large shoe dataset consisting of 50,025 shoes from a catalog on the website Zappos.com. The data was downloaded from the Zappos.com file posted on the University of Texas website in a CVS file form. The dataset is transferred to SAS language. The data is divided into four major categories: shoes, sandals, slippers, and boots. Each category is then followed by a description of the functional type (heel size, material). The dataset consists of nine categorical variables that describe the shoes. The original data included pictures. For the study, however, only the variables that describe the shoes will be used. In addition, this dataset is limited by the accuracy and completeness of online shoppers, where users pay special attention to nuanced grained visual differences (UT Zappos50K, n.d.).

Chi-square is the method chosen to analyze the data. This method has some disadvantages. One disadvantage of using this test method is that the data must be converted to numerical variables. Secondly, the test becomes invalid if any expected values are below five. Finally the third disadvantage is that all participants measured must be independent, meaning that an individual cannot fit in more than one category. However, the advantages of using the chi-square test method are that it can test the association between variables, it can identify differences between observed and expected values, and it can make no assumptions about the distribution of the population (When Chi-square Is Appropriate, n.d.).

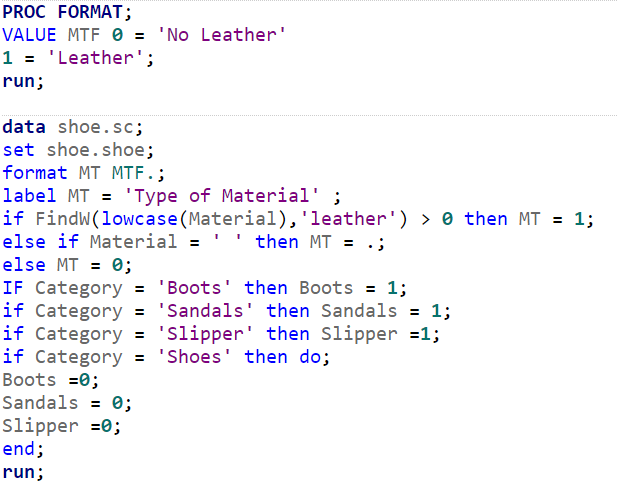
The process to import the dataset into SAS is as follows: a library was created named “shoe”, then the data was imported to SAS and stored in the new created shoe library.



Before the variable preparation, a frequency procedure was performed to find the frequency of the different categories to have a reference to create the new variables. The result shows the “Shoes” category is the most dominant among the other type of shoes.



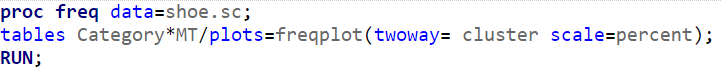
During the study, four numeric variables will be added to the dataset for shoe comparison categories. The target variable is an extraction of a “Material” categorical variable. One of the challenges of the study was to convert the categorical variables to numerical variables. Chi-square analysis requires numerical values to acquire frequencies; the categorical variables must be converted to numerical variables. However, in the case of the material “leather”, the word had to be extracted from the material variable and then assigned a number to test it. To overcome this problem, SAS has a function “FindW” which finds the word we give, in this case was “leather”. To make the function work with all the entries, first the “lowcase” function is applied, and then a low case word of leather is given. The only downside of applying these functions is that if the word is misspelled, the function will not identify the word. A format procedure is created for the new variable MT that will hold the leather’s results. The code from the format procedure and the data step is as follows:



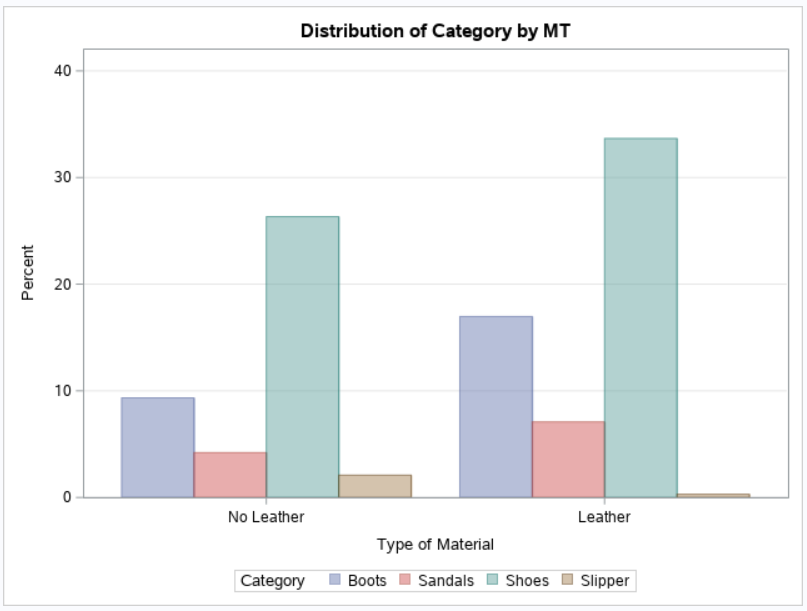
The variable preparation steps included:

* Create a format for the variable “MT” that will identify the results as 0 = “No Leather” and 1 = “Leather”.
* Extract “Leather” from the Material variable and create a numerical variable “MT”.
* Create numerical variables using an ‘if loop’ from the Shoe Categories: Boots, Sandals and Slipper identifying them as ‘1’ and Shoes as ‘0’.

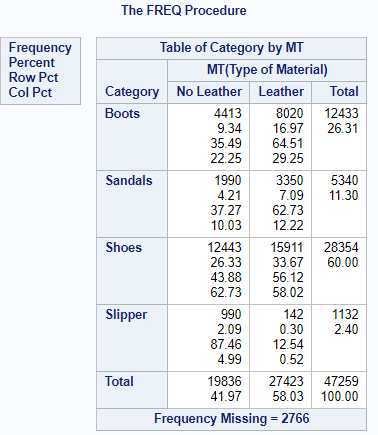
At the end of the variable preparation process, the three new numerical variables from shoe Category “Boots”, “Sandals”, and “Slipper” are compared with Leather material “MT” using the frequency procedure with the option of plots with scale in proportion and clustering the variables in groups of “Leather” and “No Leather”. The result shows a significant difference between the four categories of shoes that were manufactured with leather material. Shoes are the dominant, and Slippers are the less category of shoe manufactured with leather. The code to provide the frequency plot is:



The results from the code above are as follows:

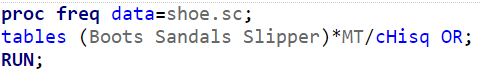


After making conclusions from the plot, another frequency procedure is performed to create different tables with more detailed results from the comparison of categories and the leather material of the shoes. Next, the following frequency table produces the proportions of each category by leather where the percentages of each category that has leather are: 64.51% boots, 62.73% sandals, 56.12% are shoes, and 12.54% slipper. Thus, overall the items that are bought online and contain leather are: 29.25% of boots, 12.22% sandals, 58.02% shoes, and 0.52% slippers. The frequency plot concludes that shoes are the best selling item manufactured with leather, and slippers the least selling item manufactured with leather.

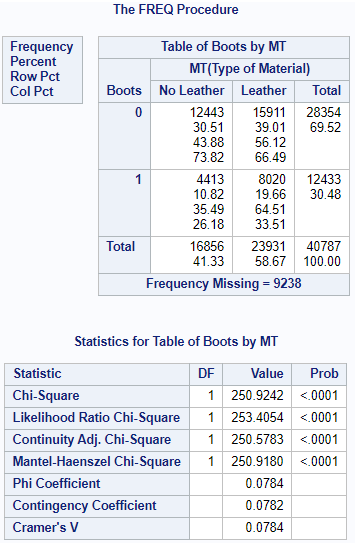


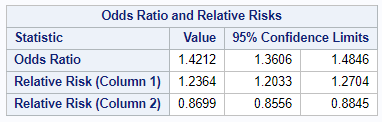
A new frequency procedure will produce the chi-square and OR (OR will give the odds ratio results and relative risk at 95% Confidence Interval) on the table’s statement using the numerical variables.

For the chi-square analysis on all the categories code:

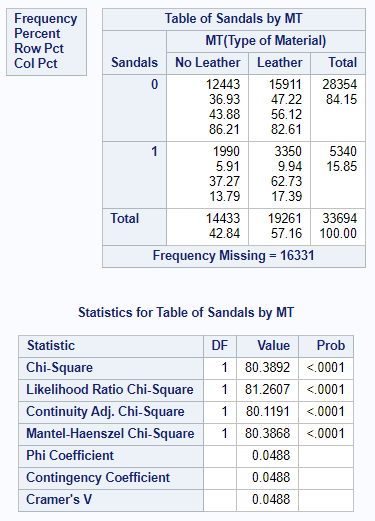


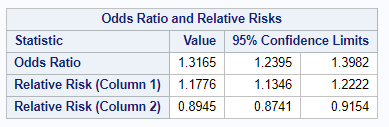
The result of the frequency table comparison with the “Boots” by “MT” against “Shoes” shows that there is a significant relationship between the boots bought online and the leather material the boots are made of, with a p-value less than 0.0001. Furthermore, they show that leather boots sell 64.51% online vs. shoes with leather of 56.12%. Thus, the odds ratio of a person buying a boot manufactured with leather is 1.4212 over buying a leather shoe (Chi-square, n.d).



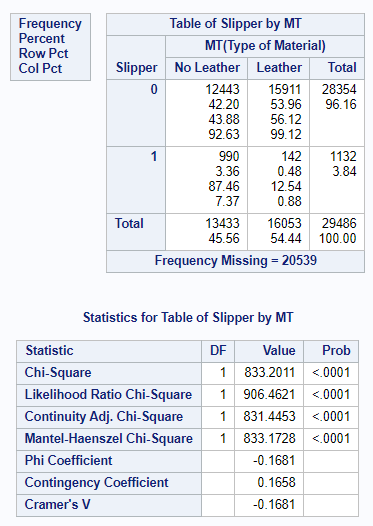


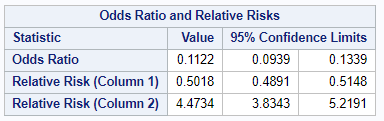
The result of the frequency table comparison with the “Sandals” by “MT” against “Shoes” shows that there is a significant relationship between the sandals bought online and the leather material the sandals are made of, with a p-value less than 0.0001. They show that sandals with leather are selling 62.73% online vs. shoes with leather of 56.12%. Thus, the odds ratio of a person buying sandals manufactured with leather is 1.3165 over buying a leather shoe (Chi-square, n.d.).





The result of the frequency table comparison with the “Slipper” by “MT” against “Shoes” shows that there is a significant relationship between the slipper bought online and the leather material the slippers are made of, with a p-value less than 0.0001. They show that slippers with leather are selling 12.54% online vs. shoes with leather at 56.12%. Therefore, the odds ratio of a person buying a boot manufactured with leather is 0.1122. Since it is less than one, we can conclude that it is less probable to buy leather slippers over leather shoes (Chi-square, n.d.).





The dataset is delimited by using the description of the shoes and does not include all the shoes’ pictures. This dataset is limited by the accuracy and completeness of online shoppers, where users pay special attention to nuanced grained visual differences (UT Zappos50K, n.d.).

 Some suggestions for future research include but are not limited to the following:

* Analyze with different manufacturer materials based on the needs of the company.
* Use different methods to accommodate the rest of the variables.
* Use the shoe images to create a model.
* Use Logistic regression analysis to find the probabilities of a material or shoe description would be preferred.

The online Shoe retailer Zappos is identified as the one with the best customer service in the world. The objective of this business, which began as a shoe shop, has been to provide a high positive customer service experience to the user. Zappos started in 1999 by investing $500,000. Despite the risks, Zappos managed to have a profit of no less than $32 million at the end of the year. In 2004, Zappos raised its profits of $184 million. In 2009, Amazon bought the company for $1.2 million with one condition: Zappos would continue cooperating as an independent brand, with its culture not being absorbed by the Amazon brand facing the customer (Ordoñez, 2018).

The findings of this study are to provide a way to find out which shoe item is manufactured with a specific material, in this case, leather. Then, in future studies, depending on the company’s needs, there could be a different material to analyze replacing leather. This way, the company can cut costs, find the most valuable materials, find the less used materials, or find less expensive material substitutes for leather shoes that do not sell well.

References:

SAS - Chi Square - Tutorialspoint. (n.d.). Www.tutorialspoint.com. Retrieved July 29, 2021, from <https://www.tutorialspoint.com/sas/sas_chi_square.htm>

UT Zappos50K. (n.d.). Vision.cs.utexas.edu. Retrieved July 22, 2021, from <http://vision.cs.utexas.edu/projects/finegrained/utzap50k>

Laia Ordoñez. (2018). Succcessful eCommerce case: the history of Zappos. Oleoshop. <https://www.oleoshop.com/en/blog/succcessful-ecommerce-case-the-history-of-zappos>

When Chi-square Is Appropriate - Strengths/Weaknesses | Chi-square Test for Goodness of Fit in a Plant Breeding Example - passel. (n.d.).

Passel2.Unl.edu. <https://passel2.unl.edu/view/lesson/9beaa382bf7e/14>

‌