Good example of statistics project essay

Finance, Banking

In 2008, there was a financial crisis which affected most of us. Till now, some countries are still recovering, including the United States. The crisis led to higher unemployment rate, lower disposable incomes, and bad loans. In all these cases, the number of loans applied by the citizens of a country and approved by lending institutions is anticipated to be lower. This will be true until a time the borrowers have increasing confidence in paying off their debt if they apply for one, and the lenders have confidence that the loans will get paid back. This will be the case when they are employed again, or companies have started to pay the employees a higher salary, or when their disposable incomes are getting higher. In this regard, the amount of loans given out by banks, financial institutions or companies might be related to the state of recovery of the economy of a country.

The data we will be using to study the health of the economy of the US is by LendingClub. com (Lending Club Statistics, n. d.). This data pertains to loans issued by the company from 2012 till 2014, from January till June. We do not analyse other months as the data for 2014 is only until June. In order to remove any cyclical effects (monthly variations in loan amounts), we choose to standardize all data in used for 2012 and 2013 to be from January till June, similar to 2014. The statistical software used here is StatCrunch (StatCrunch, n. d.).

As we have hear or read that the US is recovering every year since 2008, we shall examine if this is the case based on this data. Let the mean loan amount in dollars issued in

the i-th year be Ai (from January till June) we shall test with the following

hypotheses:

Null hypothesis, H0 : A2012= A2013= A2014

Alternate hypothesis, H1 : A2012≠A2013≠A2014

An ANOVA analysis with an F-distribution is used here.

The null hypothesis is rejected if the F test value satisfies F < Fcritical , where Fcritical is the critical value on the F-distribution corresponding to a p-value = 0. 05.

Since there are about 100, 000 total number of loans issued per year, it is impossible to calculate any statistical quantities using them all. We choose to randomly sample 100 loans per year for use in the ANOVA. For example, for year 2013 data, this is done by first randomly choosing 100 numbers from the range of numbers that corresponds to the entry numbers in the Excel sheet downloaded from LendingClub (Lending Club Statistics, n. d.) which has the issue date for the loan to be in 2013. Then, all details of the loan, and the borrower for the 100 chosen random entry numbers are extracted for further statistical analysis. The distribution of the loan amounts per year is shown in Figure 1.

Figure 1 : Loan amount distribution for 2012, 2013, and 2014

The mean and standard deviations for loan amount by year is given in Table 1.

The mean for each year is calculated by summing all the data for each year, and then divided by 99, since there are 100 observations, and 1 degree of freedom is used due to the data being a sample of a population. The sum of squares is defined as the sum of all (data – mean)2, which is equivalent to (standard deviation)2*(number of data points-1). The standard deviation is

calculated by manipulating algebraically the formula for the sum of squares. The sum of squares within group according to Table 1 is 99*(8892. 252 + 9043. 212 + 8240. 592). The value 99 comes from the fact that we have 100 data points minus one degree of freedom due to the data being a sample from a larger population. The total sum of squares is calculated by combining all the data from the three years into one, then, calculate the sum of squares for that combined data. The total sum of squares is calculated to be 299*(8706. 34). Hence, the sum of squares between groups is 299*8706. 342 - 99*(8892. 252 + 9043. 212 + 8240. 592).

The sum of squares between group is divided by the corresponding degrees of freedom which is 2 (number of groups – 1). The sum of squares within group is divided by the corresponding degrees of freedom which is number of observations minus number of groups, i. e. 297.

Hence, F-value = 299*8706. 34^2 - 99*8892. 252 + 9043. 212 + 8240. 592 2*29799*8892. 252+9043. 212+8240. 592

Thus, F-value = 0. 11. The critical value corresponding to p-value = 0. 05 is 3. 03 (F Distribution Calculator, n. d.). Since 0. 11 < 3. 03, the null hypothesis is not rejected. There are no differences between the mean loan amounts year-on-year. Based on this result, the health of the economy of the US did not change, if at all, in a significant way.

The median of the loan amount distributions is $13250 in 2012, $12000 in 2013, and $13650 in 2014. From the distributions shown in Figure 1, one can confirm that the median is always smaller than the mean. The data in each of the distributions does not seem to be normally distributed but is positively skewed towards lower loan amount region. This may be understandable if

most of the borrowers or the lender itself is risk-averse. In this case, the lender have less confidence in lending a large sum of money, and/or the borrower is reluctant to borrow beyond his/her capacity to pay back the loan.

Next, we would like to know if the company LendingClub is biased towards any borrowers that fall within certain income brackets. First, we work out the contingency table containing the debt-to-income ratio, and the income of the borrowers. Then, we perform a chi-square statistics, where the null hypothesis is that there is no bias, i. e. the income of the borrowers is independent of the debt-to-income ratio. We expect the null hypothesis since regardless of the income of the borrowers, LendingClub should only care about the debt-to-income ratio as a benchmark when giving out loans whereby the debt of the borrower upon receiving the loan has been normalized to his/her income. The alternate hypothesis is that there is a dependency between these two variables.

The contingency table for 2014 is shown in Table 2. Once again, we randomly sample 100 times the data of 2014 to obtain a sample with size 100. The number of columns is 7, and the number of rows is 6. Hence, the number of degrees of freedom is (7-1)(6-1) = 30.

The chi-squared value χ2 is calculated by:

χ2= O-E2 E,

where it is summed over each entry in the contingency table. Variable O is the observed counts in the contingency table. Variable E is the expected counts in the contingency table, for example, for the entry with income bracket (0, 5e+04] , and debt-to-income ratio of (5, 10] which has observed

count = 4, the expected count is (36/100)*13. The value 36 is the total number of counts in the (0, 5e+04] income bracket, and the value 13 is the total number of counts in the (5, 10] debt-to-income ratio.

The chi-squared value is calculated to be 31. 43. Thus, the test p-value corresponding to the chi-squared value of 31. 43 with 30 degrees of freedom is 0. 395 (Chi-Square Calculator, n. d.). Since 0. 395 is larger than 0. 05 significance level, we do not reject the null hypothesis that there is no bias, or in other words, the income of the borrower is independent of the debt-to-income ratio of the same borrower.

It would be interesting to run this chi-square test with the 2009 data, which is the year right after the financial crisis. Table 3 is the contingency table for 2009. The chi-squared value is calculated to be 19. 71, and the number of degrees of freedom is 16. The corresponding test p-value is 0. 234 (Chi-Square Calculator, n. d.). Since 0. 234 > 0. 05 significance level, we do not reject the null hypothesis also. This suggests that even just after the financial crisis, the company did not consider making a new criteria based on the debt-to-income ratio as a function of the potential borrower’s income. There is a possibility that they may have considered it, but it did not help their business. The other possibility is that they may have not consider creating a function, f(debt-to-income ratio, income) that would guide them in considering who are their potential customers hit after the 2008 crisis that can still afford to pay back the loan in time.

In this analysis, no evidence was found to support the claim that the US economy is recovering year-on-year, based on the loans issued by LendingClub. It is found also that the income, and the debt-to-income ratio of

the borrower in LendingClub is independent from one another in 2014 and 2009.

Works Cited

Chi-Square Calculator: Online Statistical Table. (n. d.) Star Trek Website. Retrieved from

http://stattrek. com/online-calculator/chi-square. aspxs

F Distribution Calculator: Online Statistical Table. (n. d.) Star Trek Website. Retrieved from

http://stattrek. com/online-calculator/f-distribution. aspx

Lending Club Statistics. (n. d.) LendingClub Website. Retrieved from

https://www. LendingClub. com/info/download-data. action

StatCrunch. (n. d.) StatCrunch Website. Retrieved from

http://www. statcrunch. com/