ASU W. P. Carey School of Business
Arizona State University

The Profitability of Pilgrim Bank's Customers

Team 12

Given the dataset at hand, we first used the *AVERAGE* function in Excel to find the average profitability of Pilgrim Bank's customers by finding the mean of the *9Profit* column, which is $111.50 and matches what Alan Green found. We next decided to find a 99% confidence interval for the profitability of Pilgrim Bank's customers so we could better understand the data. We did this by using the *STDEV.S* function to find the standard deviation of the profitability, $272.84, keeping in mind that this data is a sample. And because this data is a sample, we next found the standard error, 1.53, by dividing the standard deviation by the square root of the number of customers, 31,634, which we used the *COUNT* function to find, in the sample. Next, using the 99.5% z-score of 2.575 to construct our confidence interval since it's two-tailed, we found the 99% confidence interval via the formula $111.50 \pm 1.53 * 2.575$, resulting in a 99% confidence of [$107.55, $115.45]. This means we are 99% confident that the true population mean falls between $107.55 and $115.45.

The next order of business was to set up the proper hypothesis test to answer the question "were online customers of Pilgrim Bank more profitable than offline customers in 1999?". We decided to make our null hypothesis that the average online customer was no different than the average offline customer since we have no reason to believe that they are indeed different, and we made our alternative hypothesis that the average online customer was more profitable than the offline customer since we had to do that to complete the requirements of the assignment. In other words:

$$H_0: \bar{\mu}_{online\_customer\_proftability} = \bar{\mu}_{offline\_customer\_proftability}$$

$$H_a: \bar{\mu}_{online\_customer\_proftability} > \bar{\mu}_{offline\_customer\_proftability}$$

To test this hypothesis, we realized we needed to conduct a one-tailed test of the difference of two means and thus needed to find the average and standard deviation of both samples. In addition, because we are choosing to use a confidence interval of 99%, the z-score we find needs to be above 2.33 for us to reject the null hypothesis.

First, we used the *AVERAGEIF* function to find the average customer profitability when *9online* equaled "1" - *=AVERAGEIF(C:C,"=1",B:B)* - for online customers and when it equaled "0" - *=AVERAGEIF(C:C,"=0",B:B)* - for offline customers, resulting in an average of $116.67 for online customers and $110.79 for offline customers, both of which match what Alan Green found. We next used *COUNTIF* to find how many online customers there were via

=COUNTIF(C:C,"=1"), resulting in 3,854, and *=COUNTIF(C:C,"=0")* for the offline customers, resulting in 27,780. These two counts summed to 31,634, which is the total amount of data points we have and serves as a good sanity check.

The next step was to find the standard deviations of the profitability of the two samples. We sorted the *9Online* column from smallest to largest and, knowing there were 27,780 rows of offline customers thanks to the *COUNTIF* earlier, used *=STDEV.S(B2:B27781)* (offsetting by one to account for the header row) to find a standard deviation of $271.30 for offline customers, and used *=STDEV.S(B27782:B31635)* to find a standard deviation of $283.66 for the online customers.

Finally, we plugged the numbers we found into the following formula to find a z-score:

$$z = \frac{\overline{x_1} - \overline{x_2} - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Which gave us the following formula and answer:

$$z = \frac{116.67 - 110.79 - 0}{\sqrt{\frac{283.66^2}{3854} + \frac{271.30^2}{27780}}} = \frac{5.88}{4.85} = 1.21$$

As the z-score we found of 1.21 is lower than our z_α of 2.33, we fail to reject the null hypothesis that the two means are equal, and therefore cannot conclude with 99% confidence that the online customers are more profitable than offline customers.

To confirm this finding, we also approached by doing a single factor ANOVA test in Excel and receiving the following results:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Offline	27780	3077642	110.7862	73604.22
Online	3854	449634	116.6668	80465.63

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	117039.3	1	117039.3	1.572264	0.209888	3.841753
Within Groups	2.35E+09	31632	74439.99			
Total	2.35E+09	31633				

Since the p-value is ~0.21, we confirm that we should indeed fail to reject the null hypothesis of the means of the two groups being statistically significantly different at a 99% level.