**MSBA 620 – Statistical Foundations of Business Analytics – 1st Assignment.**

1a. For the answer to this question, we would perform a 1 sample t-test twice for each of the columns of data to ascertain the estimate mean value of each of them, the estimate means for the number of physical books read in the last 12 months and the estimate mean for the number of e-books read in the last 12 months.

**Physical Books**

Using the one-sample t-test excel worksheet/calculator made available in blackboard, I arrived at these answers:

Interval estimate: 10.227 ± 0.265

Confidence interval for the physical books: (9.910, 10.544)

Using a two-sample t-test assuming Unequal Variances I also arrive at the same answer, as shown below.



**E-books**

Using the one-sample t-test excel worksheet/calculator made available in blackboard, I arrived at these answers:

Interval estimate: 10.773 ± 0.368

Confidence interval for e-books: (10.334, 11.212)

Using a two-sample t-test assuming Unequal Variances I also arrived at the same answer, as shown below.



1b. For this question, I would be using a two-sample t-test with unequal variance. A two-sample t-test because the data I am working on is numeric and because the question requires me to measure the average of the two sets of data and find out if there is a difference or not. The initial Hypothesis or the null hypothesis is there is not difference (they are equal – there mean is equal, that on average there are equal amounts of physical book and e-books read). The parameter we are measuring is the the population mean, to determine whether the two-population mean are the same or different, there we would use the population’s sample statistics to infer about the population parameter.

­

H0: µ­­1 - µ­2 = 0 OR H0: µ­­1 = µ­2

H1: µ­­1 - µ­2 ≠ 0 OR H0: µ­­1 ≠ µ­2

I posit that the problem should be calculated as a two-sample t-test with equal or unequal variance, because I currently cannot tell if the variances of the populations are the same or not. Hence, I would perform an F-test to determine this. If the P-value from my F-test is great than (which is usually set at 0.05) we fail to reject the null hypothesis that says that the variances are equal. If the P-value from the F-test is less than (which is usually set at 0.05) then we reject the null hypothesis that the sample variances are the same, or that they have equal variances.

H0: µ­­1 = µ­2



P-value that corresponds to the test statistic: 0.04694067.

Interpret the p-value and draw a conclusion from your result.

A screenshot of a computer

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

In conclusion, there is sufficient statistical evidence to conclude that there is a difference between the number of physical books read and the number of ebooks read.

1c. For this question we would use a one-sample z-test with given value. The data we are examining is categorical. The question is basically asking us for the Lower confidence limit of the proportion of people who read 20 or more books a year (both physical books and ebooks).



P-value < **∝**

LCL, UCL: (0.785, 0.641)

Minimum value = 0.641 which is two-third of my data sample, hence we can conclude that two-third or more of the company’s customers do read twenty or more books a year.

2a. For this question we would use a matched–pair analysis test, because the data samples are related (they are data samples gathered from the same homeowners living in the same houses five years ago and now meaning there is a lot of relationship between the sample data gather 5 years ago and now, for example, the loan amount would be the same, interest rates for the house is likely to be the same, loan term is likely to be the same or similar and so many of the uncertainties that we would have in a normal two-sample t-test is already taken care of).

This is the hypothesis for the matched-pair analysis test:

H0: µ1 - µ2 = 0 OR µ1 - µ2 = µD

µD = 0

H1: µ1 - µ1 = D

H1: µD ≠ 0



In summary, we conclude that there is no difference between mortgage payments now and five years ago, our hypothesized mean (0) falls within our confidence interval of -4.09 to 28.89. Therefore, we accept the hypothesis that the mortgage payments 5 years ago and now are the same.

2b. In this question, the sample being considered are independent (as they are not related to each other), the parameter being measure are the means of the sample or population, it is two distinct samples being considered, the data type is numerical, hence we would perform a two-sample t-test with either equal or unequal variances. We would determine whether it is a two-sample t-test with equal variances or a two-sample t-test with unequal variances when we perform an F-test.

Hence, it would change the type of test I would run, instead of running a matched-pair analysis test as I did in 2a, I would run a two-sample t-test with equal variances.

The Hypothesis for our F-test is:

H0:µ1 =µ2

H1:µ1 ≠µ2



From the F-test conducted, our P-value (0.052) > ∝, therefore we fail to reject the null hypothesis and conclude that variances are the same.

The hypothesis for our T-test is:

H0: µ1 - µ2 = 0 OR µ1 = µ2

H1: µ1 - µ1 ≠ 0

H1: µ1 ≠ µ2



Conclusion: Our confidence interval is -32.71 to 57.51, ‘0’ our hypothesized means sits comfortably within this interval hence we conclude that there is not enough evidence to suggest that the average mortgage payment now is different from what it was 5 years ago.

Also looking at our p-value which is greater than ∝ we can safely fail to reject the null hypothesis which states that there is no difference between mortgage payments of this year and 5 years ago.

2c. the estimate for part (a) is 12.40 ± 16.49 and the estimate for part (b) 12.4 ± 45.11. The reason why there is a smaller margin of error in part (a) than in part (b) is because part (a) is a more controlled test, many factors of uncertainty such as size of the house, area where the house is located, bank rate, credit rate of the homeowners, etc. is controlled for in part (a) because it’s the same houseowners in the same house this year and 5 years ago, whereas in part (b) there is so much uncertainty that is not catered for.

3a. for every 1-mile increase in the odometer there is a $ -0.087 increase in y

A graph with blue dots

Description automatically generated



3b. R Square in the regression test that helps us determine the variation in the dependent variable (price), determined or explained by the odometer.

0.614 or 61.4% is the amount of variation in price that is explained by the influence of the odometer (or our model).

3c. For a used car that has 15,000 miles on it I would estimate: 15,857.51 ± 1482.43.

My confidence interval would be: (LCL, UCL) = (14375.08, 17,339.93)

LCL = 14375.08

UCL = 17,339.93

3d. Because 30,000 miles fall within the range of values on my odometer being measured, I would feel very confident using my regression equation to estimate the price of a used car that has 30,000 miles.