**ONE-SAMPLE CONFIDENCE INTERVALS & HYPOTHESIS TESTING**

**ARVIND PAWAR**

**NORTHEASTERN UNIVERSITY**

**MPS INFORMATICS**

**NUID-001444320**

**PROBABILITY THEORY AND INTRODUCTION TO STATISTICS**

**INSTRUCTOR NAME- JAMIE WARNER**

**Date- 12/02/2018**



**INTRODUCTION**

Probability Distribution is a table or equation which link each outcome of the statistical experiment with its likelihood of occurrence and possible values taken by random variables in a given range. The range would be in between the minimum and maximum possible values, but where the possible value is likely to be plotted on the probability distribution is depends on several factors. These factors include the standard deviation, skewness, distribution’s mean and so on.

In this project, we have given data of US Occupation, and I have used in-depth concepts of Probability Distribution and Sampling, that is Confidence Interval and Hypothesis Testing. Hypothesis testing is an essential procedure in inferential statistics. It is the method to test the survey or experiment to see if it has meaningful results and a theory based on inadequate evidence that lends itself to further testing and experimentation. With further testing, an assumption can usually be proven true or false. Confidence Interval is a range of values defined as there is a specified probability that the value of the parameter lies within it. In this assignment, I have worked on the excel sheet with given dataset and have implemented confidence interval and hypothesis testing.

**Part 1**

**Introduction:**

In the first part, I have asked to generate a random sample of size 40 from LOC Quotient(Population) column. I used inbuilt excel function Data Analytics which I have used in previous modules. From Data Analytics I used sampling function, and I gave LOC Quotient data (A2: A7287) as input to generate 40 random numbers at column D.

Later on, I Calculated the Population Mean using the inbuilt function AVERAGE():

**H2=AVERAGE(A2: A7287)**

The Population Mean is 1.04

Calculated the Population variance using the inbuilt function VAR.P():

**H3= VAR.P(A2: A7287)**

The Population Variance is 0.84

Calculated the Population Standard Deviation using the inbuilt function STDEV.P():

**H4= STDEV.P (A2: A7287)**

The Population Standard Deviation is 0.92

**Part 2**

In part 2, I have taken the random sample of size 40 that generated in part 1. In this part, first I have calculated Standard Deviation, Mean, Sample Size and Sampling Error of the random sample.

In **Table A,**

Calculated the Sample mean using the inbuild function AVERAGE():

F2=AVERAGE(A2:A41)

The Sample Mean= 1.01

Calculated the Sample Standard Deviation using the inbuild function STDEV.S():

F3= STDEV.S(A2:A41)

The Sample Standard Deviation = 0.49

Calculated the Sample size using the inbuild function COUNT():

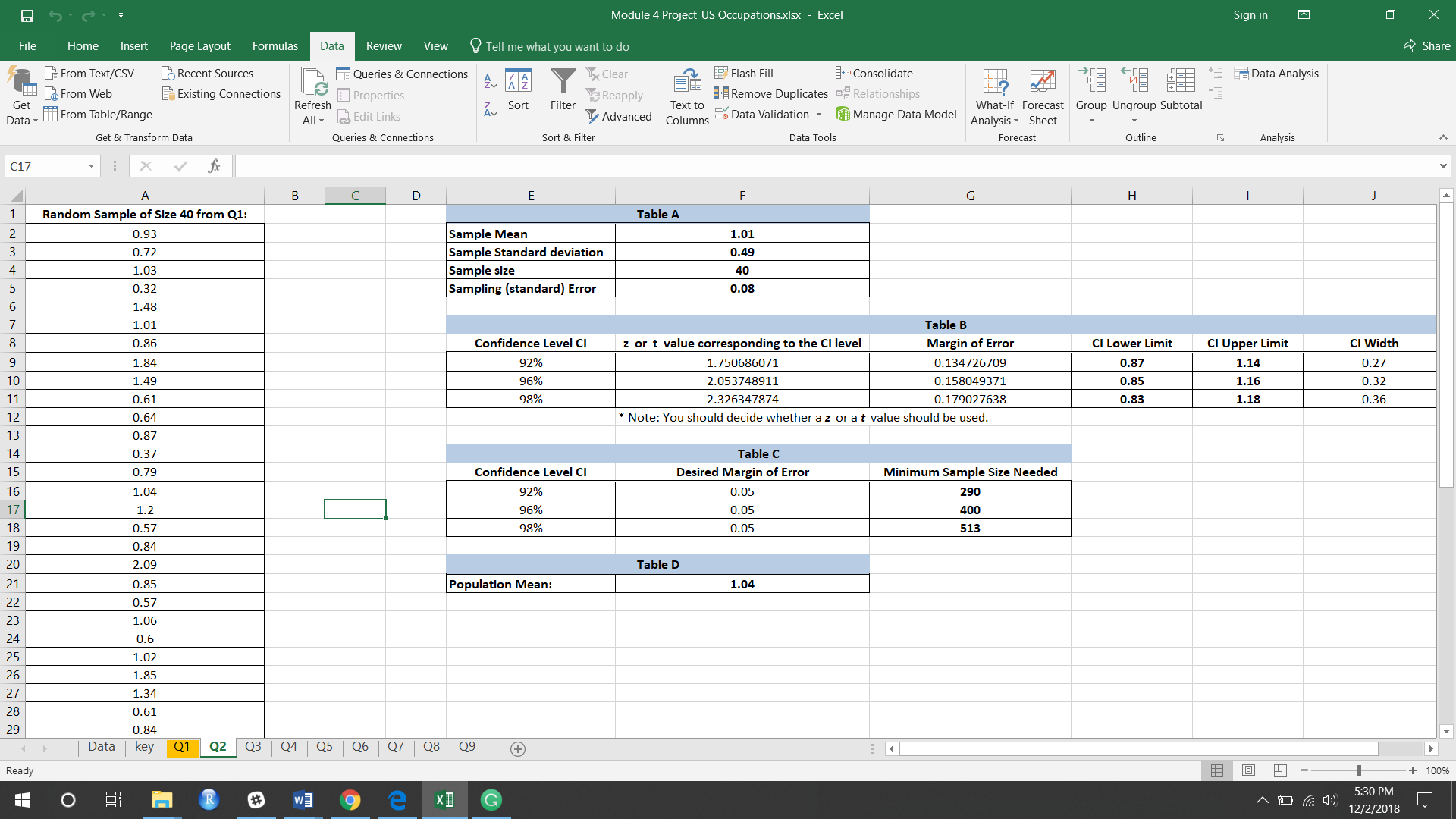
F4= COUNT(A2:A41)

The Sample size = 40

Calculated the Sample (standard) Error:

Sample (standard) Error(F5) = Standard Deviation (F3) / Sample size (F4) =0.33/50=0.05

The Sample (standard) Error= 0.08



In **Table B**,

To create 92%, 96% and 98% confidence intervals for the population mean, I used the given sample. First I need to calculate the z score for the given confidence intervals. To calculate z value, I have used **((NORM.S.INV((1+c)/2))) function** based on the formula of Zc which is as follows:

**(1 + C)/2**

**As** zc is a specific value of standard normal random variable Z corresponding to given level of confidence. So I have used **((NORM.S.INV((1+c)/2))) function.**

For 92% confidence interval I calculated **z-value** as follows:

=NORM.S.INV((1+E9)/2) = 1.750686071

**Margin of Error** = (Standard Deviation \* z-value) / √ Sample size

= (0.49\*1.750686071) / √40

= 0.134726709

**CI Lower Limit**:

Sample Mean – Margin of Error = **F2-G9** = **0.87**

**CI Upper Limit**:

=Sample Mean + Margin of error = F2+G9 = **1.14**

**CI Width**:

=Upper limit – lower limit

= I9-H9 = 0.27

Similarly, I have calculated everything for 96% and 98% confidence intervals.

In **Table C**:

Desired Margin of error is given as 0.05.

Calculated the Minimum sample size needed:

=((Standard Deviation \* z-value) / Desired Margin of error)^2

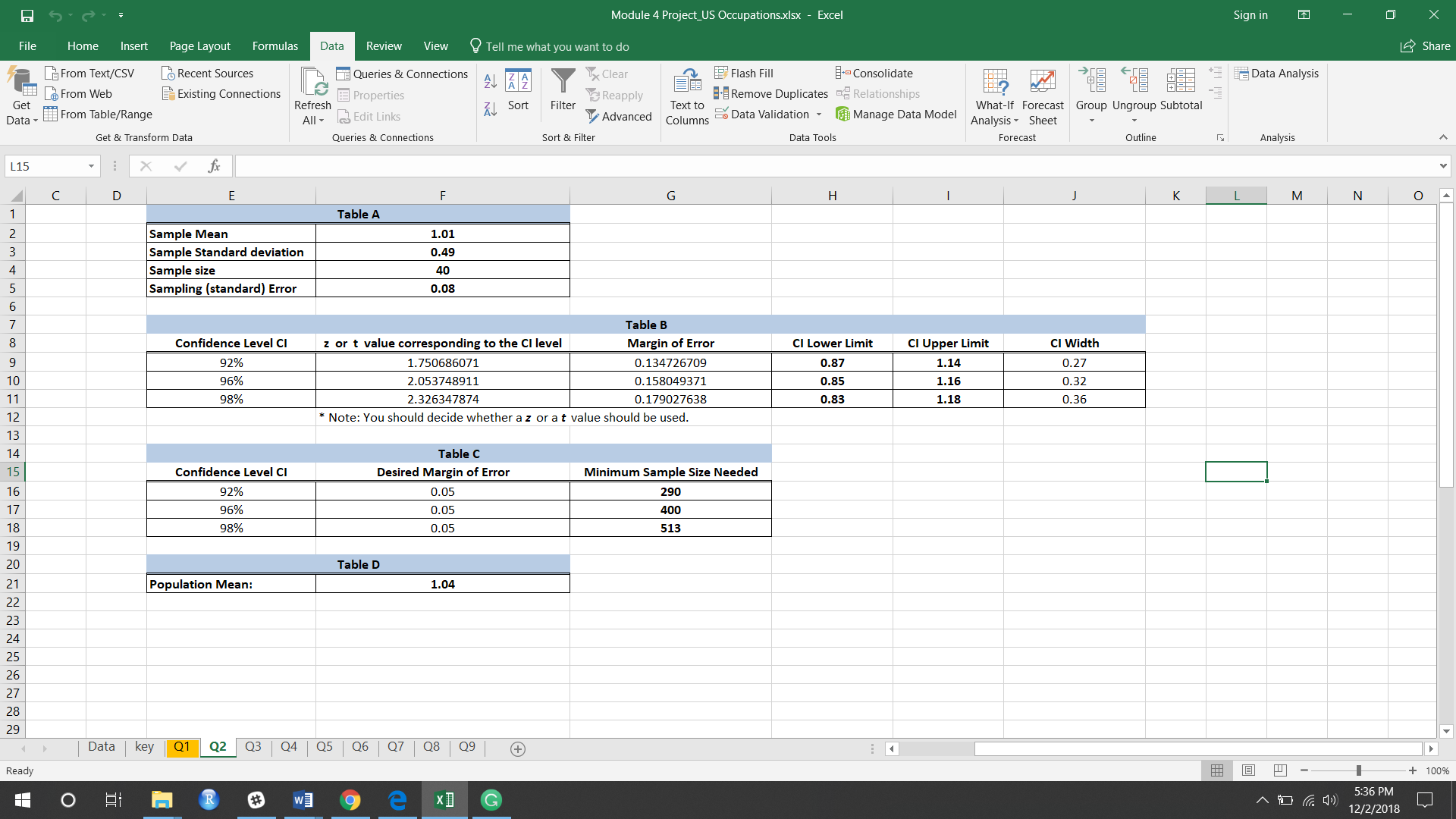
= (0.49 \*1.750686071) /0.05

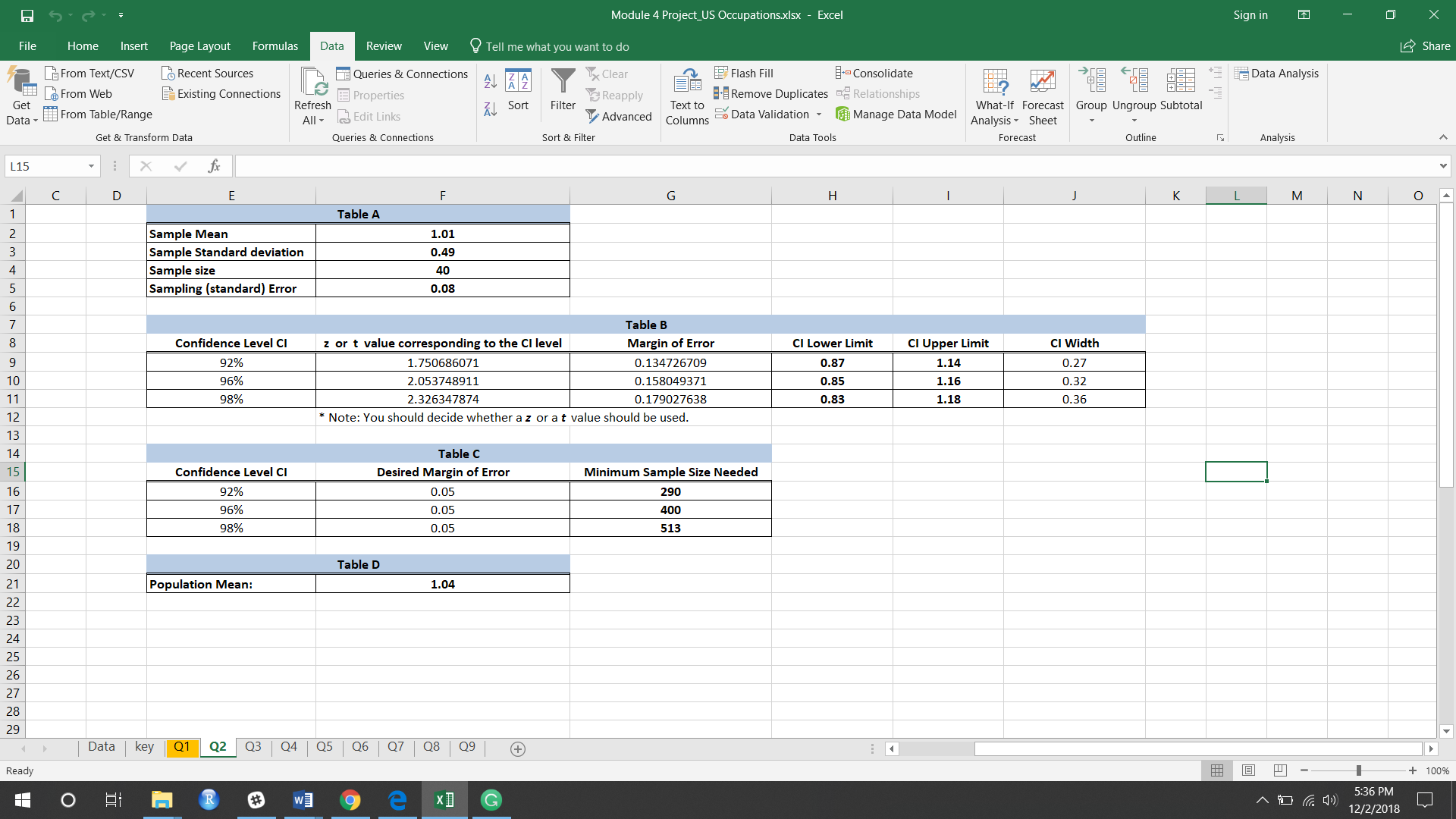
Minimum Sample size with CI 92% = **290 (I rounded the value to next whole number)**

Similarly, for

96% = **400**

98% = **513**





**Analysis:** All three-confidence interval contain the population mean **1.04.**

**Part 3**

**In Part 3, I am using the same population data, I have generated the random sample of size 24.** I used inbuilt excel function Data Analytics which I have used in previous modules. From Data Analytics I used sampling function, and I gave LOC Quotient data (A2: A7287) as input to generate 24 random numbers at column F.

Later on, I Calculated the Sample Mean using the inbuilt function AVERAGE():

**H2=AVERAGE(F2:F25)**

The Sample Mean is 1.09

Calculated the Sample Standard Deviation using the inbuilt function STDEV.S():

**H3= STDEV.P (F2:F25)**

The Population Standard Deviation is 0.60

Calculated the Sample size using the inbuild function COUNT():

F4= COUNT(A2:A41)

**The Sample size = 24**

Calculating the **Sample (standard) Error**:

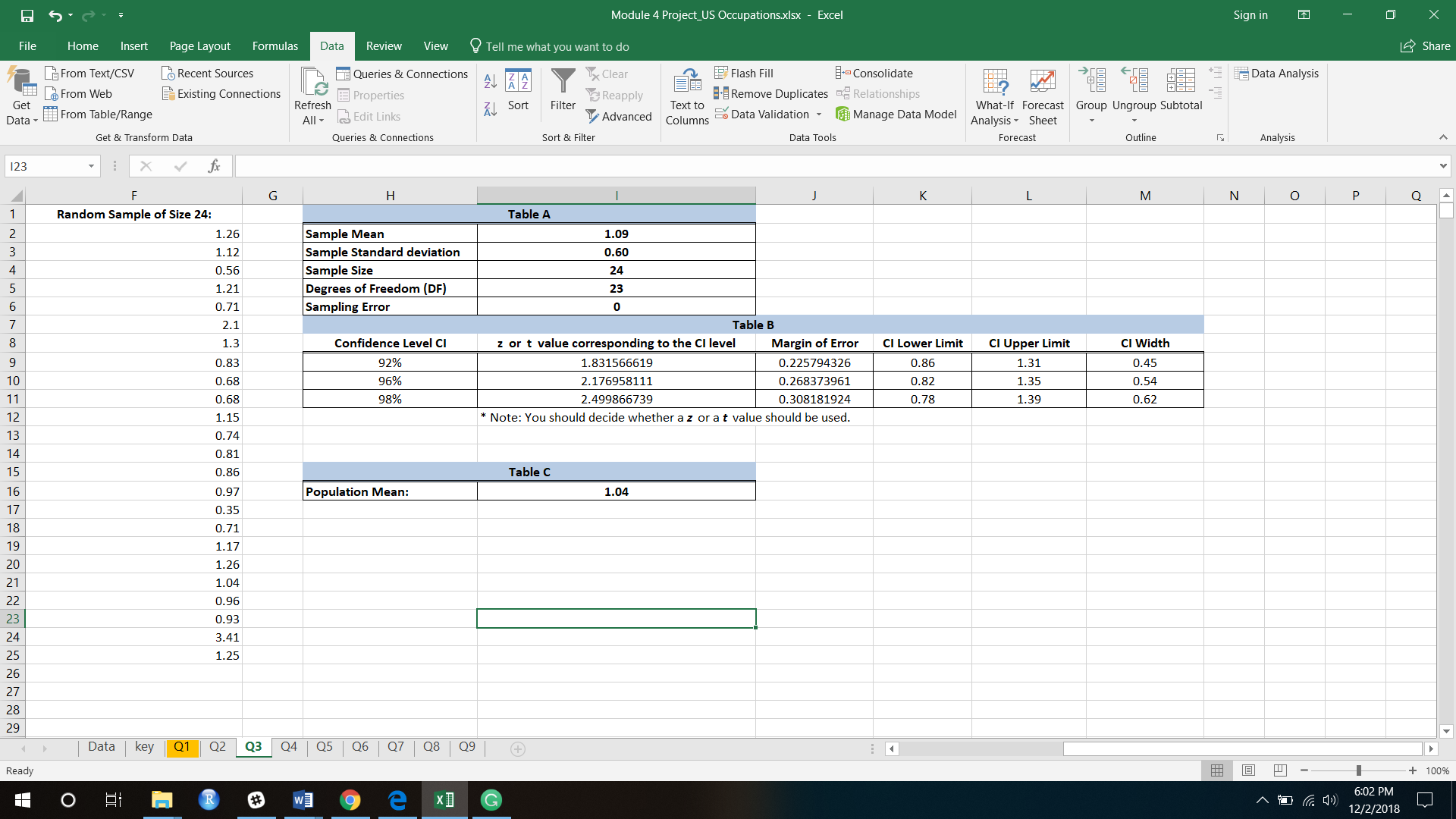
Sample (standard) Error(F5) = Standard Deviation / √Sample size (F4) =0.60/√24=0

**The Sample (standard) Error= 0**

**Degree of Freedom(DOF)**

=Sample size-1

=24-1= 23



For **Table B**, First, I need to calculate the t values for the 92%, 96% and 98% confidence intervals as the sample size is less than 30.

To calculate t-value we have used the inbuild function T.INV(Probability, DOF) which returns the left-tailed inverse of the t-distribution.

For **92% confidence interval** I calculated **t-value** as:

=T.INV(((1+0.92)/2),I5)

= 1.831566619

**Margin of Error** **= (Standard Deviation \* t-value) / √ Sample size**

= (0.60\* 1.831566619) / √24

= 0.225794326

**CI Lower Limit:**

=I2-J9= 0.86

**CI Upper Limit:**

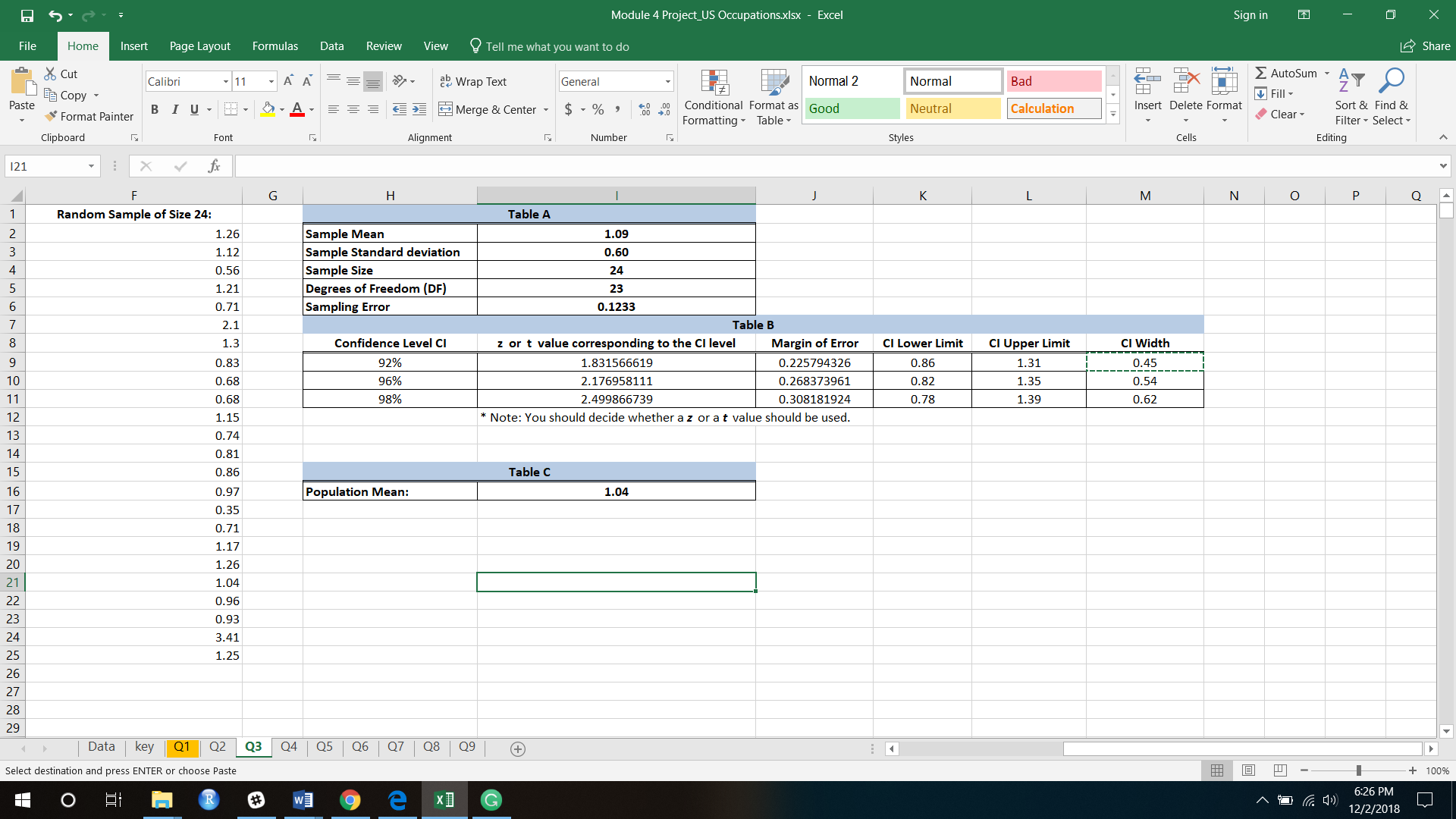
=I2+J9= 1.31

**CI Width:**

=Upper-lower

=L9-K9= 0.45

Similarly, I calculated everything for **96%** and **98%** Confidence Level.



**Analysis:** From the above table we can conclude that the **population mean 1.04** lies in all **three-confidence interval.**

**Part 4**

Even in this question I am using same population data, I have generated a random sample of size 180, I did it using excels inbuilt function Data Analysis which we have used it before in our previous modules.

In Data Analysis I used sampling function and in this I gave LOC Quotient data as input to generate 180 random sample in (C2: C181).

In **Table A**

Calculating Proportion of success using inbuild function COUNTIF()

=COUNTIF(C2:C181,"<1")/F4

**The Proportion of success =** **0.3389**

Calculating Proportion of failure

= 1-Success = 1- **0.3389**= **0.6611**

**The Proportion of Failure =** **0.6611**

Calculating the Sample size using the inbuild function COUNT():

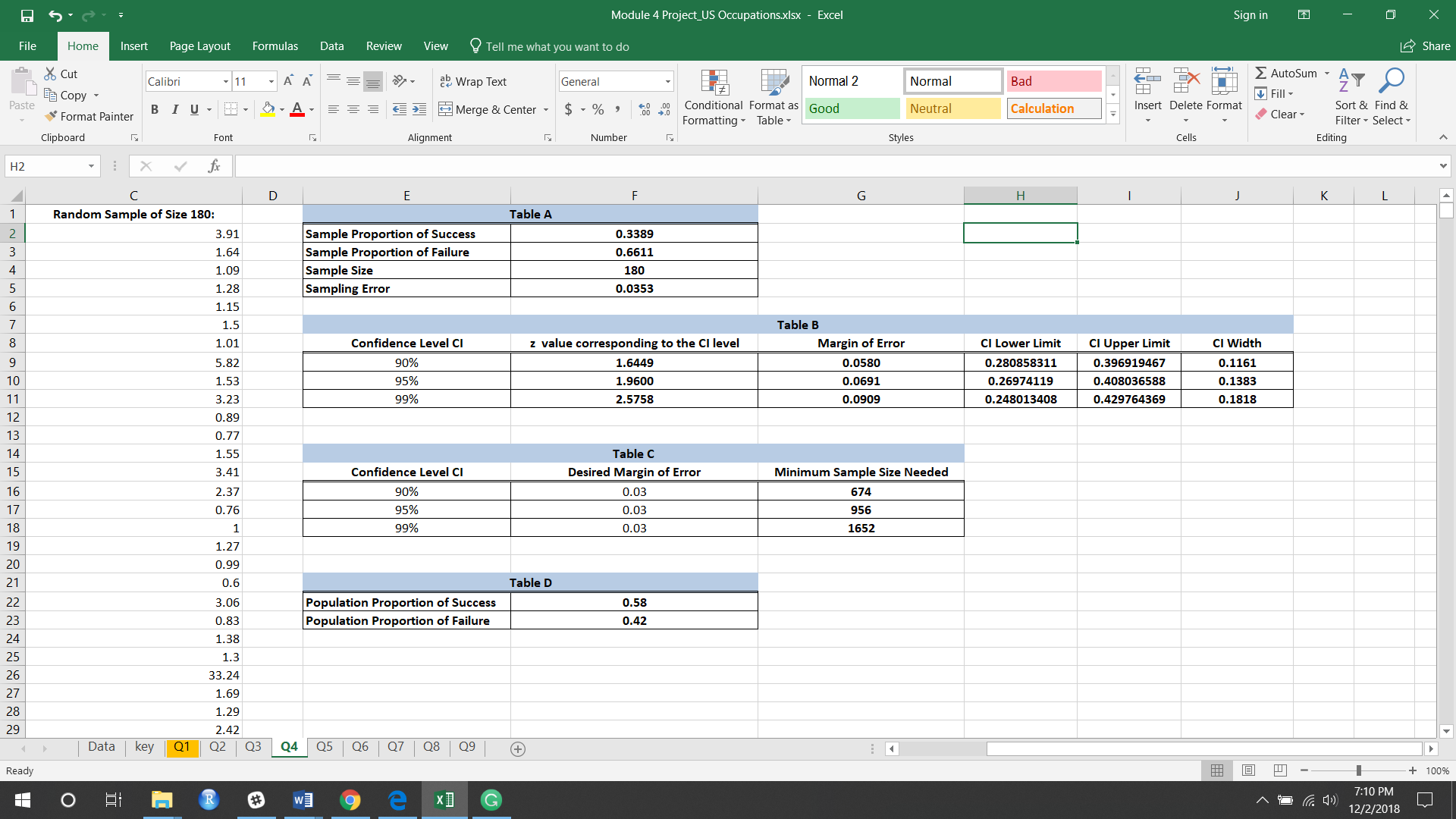
= COUNT(C2: C181)

**The Sample size= 180**

Calculating the Sampling Error:

=√(Success\* Failure)/sample size

=SQRT((F2\*F3)/F4) = **0.0353**



For Table B, First, I need to calculate the z values for the 90%, 95% and 99% confidence intervals.

To calculate z-value we have used the inbuild function NORM.S.INV() which returns the inverse of the standard normal cumulative distribution, for distribution which has a mean of 0 and a standard deviation as 1.

For 90% confidence interval we calculate **z-value as**:

=((NORM.S.INV((1+0.90)/2)))

= 1.6449 which is the z value for 90% interval

**Margin of Error** =90% CI\*(√(Success\* Failure)/sample size)

= **0.0580** gives the margin of error for 90% interval

**Confidence interval Lower Limit:**

=Sample Success-Margin of error

=F2-G9= **0.280858311**

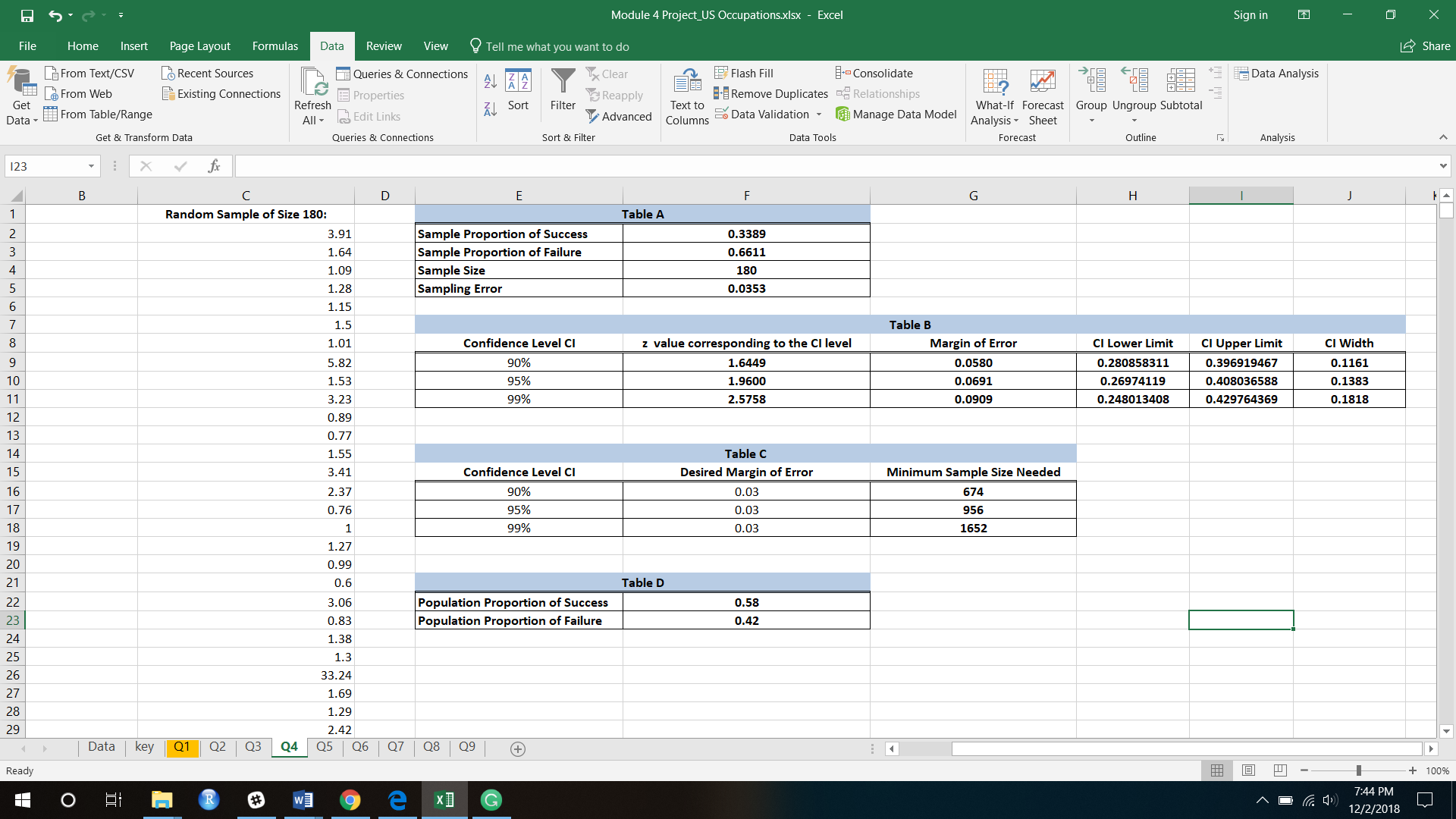
**Confidence interval Upper Limit (H9):**

=F2+G9= **0.396919467**

**CI width**

=I9-H9= **0.1161**

Similarly, I have calculated everything for 95% and 99% Confidence Level.



In Table C, Desired Margin of error is given as 0.03.

Calculating Minimum sample size needed:

=0.5\*0.5\*(H9/H16)^2

90% Minimum sample size = 752

Similarly,

95% Minimum sample size = 1067

99% Minimum sample size = 1843



**Table D**, Population Proportion of Success is calculated by this function =COUNTIF(A2:A7287,"<"&1)/COUNT(A2:A7287)

= **0.58**

Population Proportion of Failure:

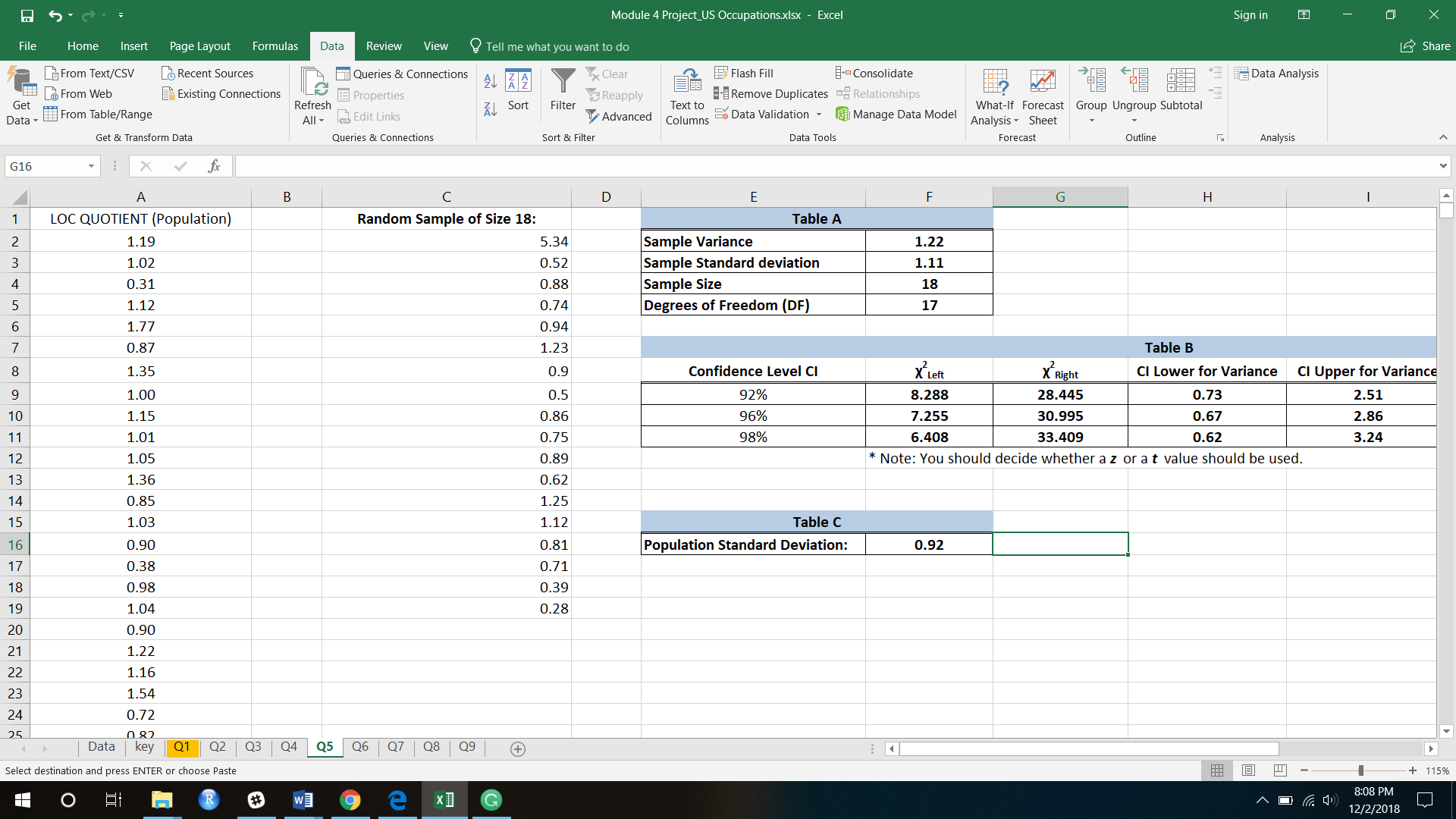
=COUNTIF(A2:A7287,">="&1)/COUNT(A2:A7287)=1-0.1672= **0.42**



**Analysis:**

**Part 5**

In Part 5, I generated the random sample of size 18 and calculated the below values by using functions which I used in previous questions.



I used sample to create 92%, 96% and 98% confidence intervals for the population  
variance and standard deviation.

For 92% CI

X^2 Left we have calculated using:

=CHISQ.INV((1-E9)/2,$F$5) F5 is DOF

= **8.288**

X^2 Right we have calculated using:

=CHISQ.INV((1+E9)/2,$F$5) F5 is DOF

= **28.445**

CI Lower for Variance:

=$F$5\*$F$2/G9 = **0.73**

CI upper for Variance

=$F$5\*$F$2/F9= **2.51**

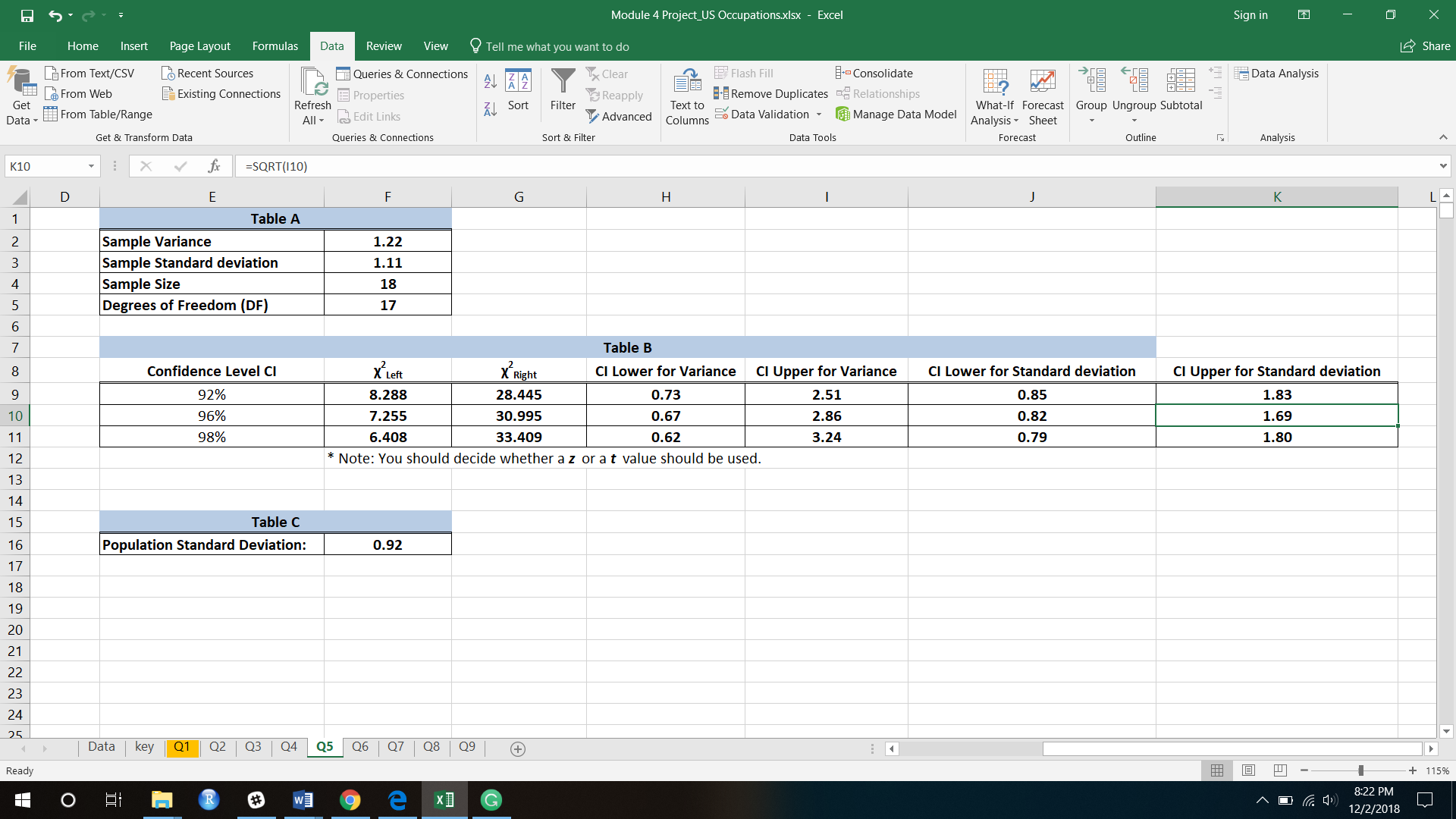
CI Lower for Standard deviation

=SQRT(H9) = √0.73 = **0.85**

CI upper for Standard deviation

=SQRT(I9) = √2.51= **1.58**

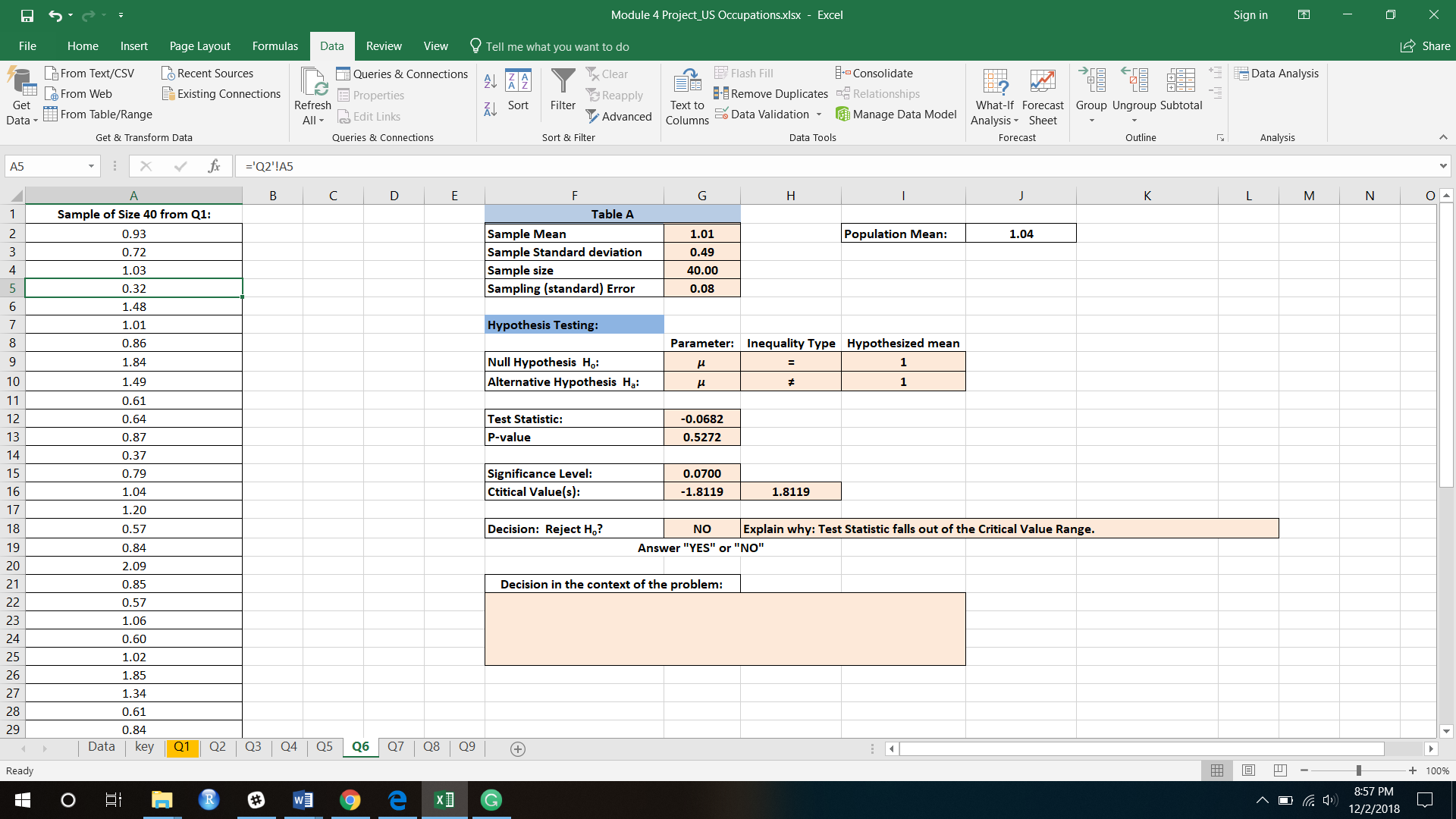
Similarly, we calculated for 96% and 98%.



**Analysis:** The Population standard deviation is 0.92, from above we can say all three-confidence interval contain the population standard deviation.

**Part 6**

**Introduction:** In part 6, I did test the hypothesis to check if the population mean is different from 1. I have calculated the Sample Mean, Sample Standard Deviation, Sample Size and Sampling error in similar way that I did in previous questions.





**Hypothesis Test:**



**As given in the question, alternative hypothesis is different from 1.**

**Calculating Test statistics:**

**(Hypothesis Mean – Sample Mean)/Sampling error**

**Test statistics =(I9-G2)/G5 = -0.0682**

**p value =1-NORM.S.DIST(G12,TRUE) = 0.5272**

**Critical value range is calculated using;**

**=NORM.S.INV(G15/2 )= 1.8808**

**=NORM.S.INV(1-G15/2) = -1.8808**





**Analysis:** The significance level is given in the question. α, is the probability of the study rejecting the null hypothesis, given that it was true. This is a two-tail test, we have to consider both the sides from mean in our normal distribution. Test Statistic is the summary of the information observed in the sample. The Test Statistic value is within the Critical Value Range. So, accepting the hypothesis that the population mean is different from 1.



**Conclusion:**

I see that p value is 0.5272 that means the probability of mean of population to not be equal to 1 is 53%, which is not breaking the threshold required, which is 3.5% on any side of normal distribution. if the p value would lie in that .07 significance level then we can reject the null hypothesis towards the alternative hypothesis.

**Part 7**

**Introduction: In part 7, I did test the hypothesis to check the population mean is greater than 1. In this step I have performed hypothesis test on sample with size 24 which is less than that of previous question. I have calculated the Sample Mean, Sample Standard Deviation, Sample Size and Sampling error, Test Statistics, Critical Value, p-value in similar way that I did in previous questions.**











**Analysis:**

As the sample is less the 30, it will not make a normal distribution graph. So, I have used different formula to calculated z values accordingly. Also, it is a right tailed test because the test is to figure out population mean is greater than 1. Because 1 is the measuring number, in our null hypothesis, we assume it to be the mean of our normal distribution. Hence the z value tells us as to how far our sample mean is from the assumed number of hypothesis with respect of standard deviation.



**Conclusion:**

The Probability of null hypothesis does not lie beyond the significance level. The Test Statistics value is less than Critical Value. Also, we have considered that is left tail test as we are doing the test for value greater than 1. So, I failed to reject this hypothesis. In other words, if we go from left to right on our normal distribution chart, we see that the mean of our population can be less than 1 with only 67% surety. If the surety would have been 93% or more, we could have rejected it since the required significance level if set at 7%.

**Part 8**













**Analysis:**

In this step I have sample data of size 180, which is greater than 30. We can say that the distribution is a normal distribution. It is a left tail test as we are testing our hypothesis if it is less than 65%, hence left side of the mean. We can see that the value of z is negative, which means it is on the left side of the 0.

**Conclusion:** I failed to reject because we do not have the enough confidence to reject the null hypothesis. As our limit is set at 1% from left to right on normal distribution, the demanded critical values is -2.32, which means 2.32 left of the mean, but our z-value of sample is not on extreme left of the mean on distribution. The population proportion in Q4 is 58% and here we are comparing it with 65%. Population proportion is less than hypothesis proportion. So I failed to reject hypothesis and my observation is correct.



**Part 9**













**Analysis:**

This question has two tailed test, because greater than or less than condition is not given. This means we will compare our hypothesis at both ends of normal distribution. Moreover, here I am going to compare standard deviations.

**Conclusion:**

I failed to reject this hypothesis because my p value is less than my significant level which is set at 1%. The test statistic value lies between the critical values.

Bibliography:

* Sampling Distribution of the Sample Proportion, p-hat. (n.d.). Retrieved from <https://bolt.mph.ufl.edu/6050-6052/module-9/sampling-distribution-of-p-hat/>
* Confidence interval. (2018, November 24). Retrieved from <https://en.wikipedia.org/wiki/Confidence_interval>
* <http://www.mit.edu/~6.s085/notes/lecture2.pdf>
* Sampling Distribution of the Sample Proportion, p-hat. (n.d.). Retrieved from <https://bolt.mph.ufl.edu/6050-6052/module-9/sampling-distribution-of-p-hat/>
* Stat Trek. (n.d.). Retrieved from <https://stattrek.com/statistics/dictionary.aspx?definition=probability_distribution>
* <https://northeastern.acrobatiq.com/courseware/NEU_MPSA_PROB_THEORY_INTRO_STATS_ALY6010_V2_2/week_4_-_confidence_intervals_and_hypothesis_testing_one_sample/module_overview__confidence_intervals_and_hypothesis_testing_one_sample/wbp_confidence_interval_ci_for_a_population_mean>