

Title: Confidence and Hypothesis

Introduction: The purpose of this assignment is to take sets of data corresponding to different demographics: Gender, Company sizes, Salaries, and Savings, to determine how confident we can be about the probabilities amongst variables. Specifically working with calculating confidence levels of means. Along with determining whether or not to reject null hypothesis' by evaluating by methods of p-values and critical values. All the data presented here has been analyzed and processed through Microsoft Excel.

Table of contents:

- ❖ Confidence Intervals Data by Sector ~ Pages 2 - 4
- ❖ Confidence Intervals Data by Company Size ~ Pages 5
- ❖ Hypothesis & Claims with medium size companies as population ~ Pages 6 - 7
- ❖ Hypothesis & Claims with entire population data ~ Pages 8 - 9
- ❖ Conclusion ~ Page 10
- ❖ References ~ Page 11

1) Sort your data by Sector.

- a. **Stratify your data by Sector and randomly select 5 rows from each stratum.**
- b. **Copy your sample in a new spreadsheet and work with this.**
- c. **Construct a 95% confidence interval for the mean of the salaries.**

The confidence level was given. The alpha was 0.05 since $1 - 0.95 = 0.05$. After selecting 5 rows from each stratum I calculated the mean of that sample. Since the confidence level was 0.95, I was able to determine the confidence coefficient was 1.96 from the z-table. The sample size was calculated to be 30 since there were 5 samples per each sector and there were 6 sectors in total. The standard deviation was calculated via excel. The standard error was calculated by taking the standard deviation divided by $\sqrt{\text{sample size}}$. To calculate the confidence interval, we had to find the High End = sample mean + $z\text{-score}(\alpha) * s/\sqrt{n} = 84831.912$ and the Low End = sample mean - $z\text{-score}(\alpha) * s/\sqrt{n} = 68719.708$.

Therefore, the confidence interval for the mean is: $68719.708 < \mu < 84831.912$

Confidence level	0.95	
Alpha	0.05	
All Sample Mean	76775.81	
Confidence coefficient	1.96	
Sample size	30	
Overall Standard Dev	22512.8	
Standard Error	4110.2561	
Confidence Interval		
68719.708	μ	84831.912

d. Construct a 99% confidence interval for the mean of the salaries.

As in part **c**, the same steps are taken to find the mean confidence interval for a 99% confidence level, where alpha is now 0.01.

Resulting in and confidence level of: $66187.7902 < \mu < 87363.8298$

Confidence level	0.99	
Alpha	0.01	
All Sample Mean	76775.81	
Confidence coefficient	2.576	
Sample size	30	
Overall Standard Dev	22512.8	
Confidence Interval		
66187.7902	μ	87363.8298

e. Compare the two confidence intervals and explain their differences based on the theoretical concepts involved.

The interval for a 95% confidence level resulted in $68719.708 < \mu < 84831.912$

While the interval for 99% a confidence level resulted in $66187.7902 < \mu < 87363.8298$.

It is apparent that there is a 2531.9178 difference between both Low Ends and a -2531.9178 difference between the High Ends. By examining the interval differences, we see that the intervals have the same magnitude while starting and ending in different points. The differences of the starting

points and ending points are due to the percent of confidence we are testing for. Since we are going from a 95% to a 99% confidence the interval magnitude remains the same, while there is a shift of the interval to the left since we are including a larger percentage.

- f. **Calculate the population mean of salaries and verify if your sample is a good sample. If it isn't repeat the process, until you find a good sample.**

The population mean was calculated to be:

Population mean	82764.68
-----------------	----------

The population mean fell between both intervals: $68719.708 < \mu < 84831.912$ and

$66187.7902 < \mu < 87363.8298$. The fact that the population mean fell between the two mean sample intervals verifies that the samples that were taken was a good sample, because it gave an accurate estimate to where the population mean may fall upon.

2) Sort your data by Company Size

- Stratify your data by Size and randomly select 3 rows from each stratum.
- Copy your sample in a new spreadsheet and work with this.
- Construct a 95% confidence interval for the proportion of salaries greater than 40,000\$

As in part 1.c we evaluated what alpha was, calculated the sample mean, the confidence coefficient, the sample size and the population standard deviation. Since we are working with proportions we need to ensure the $np > 5$ and that $nq > 5$. In my sample data I calculated that 11/12 ($=p$) of people made a salary greater than \$40,00 while 1/12 ($=q$) made less than that.

It was calculated that $np = 11$ and that $nq = 1$, due to the fact that nq was less than 5, determining a confidence interval is not possible.

Confidence level	0.95		n	12
Alpha	0.05		p	0.916667
All Sample Mean	52044.37		q	0.083333
Confidence Coff	1.96		np	11
Sample size	12		nq	1
Overall Standard Dev	17657.62			

- Calculate the population proportion of salaries $> 40,000$ \$ and conclude. No need to repeat if the sample is not good.

The population proportion of salaries that makes over \$40,000 was \$86130.16. This comes from taking the mean of those who make more than \$40,000, meaning that on “average” those who make over \$40,000 make around \$86130.16.

Population > 40 Mean salary	86130.16
-------------------------------	----------

3) Consider the entire data set and consider the medium size companies only as your population.

- a. Consider the top 32 rows of your sample as a randomly selected sample of the entire population.
- b. Run a hypothesis test for the Claim, the savings of individuals working in medium size companies is on average less than 17,000\$
- c. Find the H_0 and H_a

H_0 : salary greater than or equal to 17000		
H_a : salary less than 17000	Claim	Right-Tailed Test

- d. Consider a significance level 0.01 for your test.
- e. Run the appropriate test with both methods (p-value and critical value).
- f. Show every one of the steps and explain your work, sharing formulas used, rationale of choices, final answers, conclusions.

In this particular problem I started by finding the sample mean for savings and finding the sample standard deviation. I then solved for the population mean for savings and the population standard deviation. All calculated via excel. The sample size was given to be 32 along with the significance level which was 0.01. I calculated the degree of freedom $= n-1 = 31$. This was enough information to solve for the t-table critical value. After finding that the critical value was 2.453, I evaluated for the test-statistic to compare it.

Using the formula for the T-Stat $= \text{sample mean} - \mu / \text{sample standard deviation} / \sqrt{n}$

$= -0.779348304$, because this value is less than 2.453 we reject the H_0 . For the p-test we use z-scores.

We must calculate the area for z, we do this by taking $1-\alpha/2 = (1-0.01)/2 = 0.49$. After finding that the area is 0.49 we find the z-score from the table, which equals 0.03. We can now take the test-statistic to compare. The formula for the z statistic $= \text{sample mean} - \mu / \text{population standard deviation} / \sqrt{n} = -0.779348304$. Since this value is less than the z-score of 0.03 we reject the H_0 .

Rejecting H_0 here means that the savings of individuals working in medium size companies is on average not greater than or equal to \$17,000.

Sample Mean Savings	52636.90625	
Sample Standard Dev	105065.6422	
Population mean savings	67111.86466	
Population Standard Dev	119826.7886	
Sample size	32	
Alpha	0.01	
Degree of Freedom	31	
test-statistic (USE WHEN estimate for STDEV)	-0.779348304	Less than 2.453 so reject H_0
t-table critical value	2.453	
area for z (alpha=0.01, take 1-a/2)	0.49	
z-score	0.03	
z-stat	-0.779348304	Less than 0.03 so reject H_0

4) Assume the Claim: Men and Women make the same amount of money in the entire population of the 500 persons. Consider all those working in Large Size companies as your sample and test this claim with $\alpha=0.05$.

- a. Show every one of the steps and explain your work, sharing formulas used, rationale of choices, final answers, conclusions.

Ho: salary Men = salary Women	claim	two tailed test
Ha: salary Men not = salary Women		

I began by finding the sample mean for savings and finding the sample standard deviation. I then solved for the population mean for savings and the population standard deviation. All calculated via excel. The sample size was given to be 184 along with the significance level which was 0.05. I calculated the degree of freedom = $n-1 = 183$. This was enough information to solve for the t-table critical value. After finding that the critical value was 1.972, I evaluated for the test-statistic to compare it.

Using the formula for the T-Stat = $\text{sample mean} - \mu / \text{sample standard deviation} / \sqrt{n}$
= -2.32162687, because this value is less than 1.972 we reject the Ho. For the p-test we use z-scores. We must calculate the area for z, we do this by taking $1-\alpha/2 = (1-0.05)/2 = 0.45$. After finding that the area is 0.45 we find the z-score from the table, which equals 1.645. We can now take the test-statistic to compare. The formula for the z-statistic = $\text{sample mean} - \mu / \text{population standard deviation} / \sqrt{n} = -2.32162687$. Since this value is less than the z-score 1.645 we reject the Ho. Rejecting Ho here means that we reject the hypothesis that the salaries between both sexes are equivalent.

Sample Mean Salary	77896.58336	
Sample Standard Dev	28443.04558	
Population mean salary	82764.68466	
Population Standard Dev	29036.31	
Sample size	184	
Alpha	0.05	
Degree of Freedom	183	
test-statistic (USE WHEN estimate for STDEV)	-2.32162687	Reject Ho, less than -1.972
t-table critical value	1.972	
area for z	0.45	
z-score (alpha=0.05, take 1-a/2)	1.645	
z-stat	-2.32162687	Reject Ho, less than -1.972

Conclusion:

The study of this given data was successful, we were able to make many conclusions based of evidence. In Part 1 we saw how 2 different confidence intervals for the mean salaries compare and differ. In Part 2 we showed that it was not possible to construct a confidence interval for the proportions of salaries because $nq < 5$. In Part 3 and 4 we ran hypothesis test for claims using the appropriate test with both method of p-values and critical values. What might be learned from conducting these types of calculations and test is how to make decisions or develop certain traits based off where one fall on a given interval.

References:

Confidence Interval Basics – dummies Retrieved 5 May 2018 from

<http://www.dummies.com/education/science/biology/confidence-interval-basics/>

Pierce, Rod. (9 Apr 2018). "Confidence Intervals". Math Is Fun. Retrieved 5 May 2018 from

<http://www.mathsisfun.com/data/confidence-interval.html>

Pierce, Rod. (10 Apr 2018). "Confidence Interval Calculator". Math Is Fun. Retrieved 5 May 2018

from <http://www.mathsisfun.com/data/confidence-interval-calculator.html>