Course Project - Phase 5 Final Submission

Shaun Pritchard

Rasmussen College

STA3215

Fie Wang

May 10, 2020

**Couse Project Index**

[Course Project – Phase 1](#_Course_Project_–) 3

[Course Project – Phase 2](#_Course_Project_–_1) 7

[Course Project – Phase 3](#_Course_Project_–_2) 14, 21

[Course Project – Phase 4](#_Course_Project_–_3) 21

[References](#_References) 22

*Course Project - Phase 5 Final Submission*

*There where no remarks or corrections to be made for the submission of this final proof.*

# Course Project – Phase 1

**Introduce your scenario and data set.**

Provide a brief overview of the scenario you are given above and the data set that you will be analyzing.

The data set is based on a new infectious disease in which will determine if the age of patients correlates or facilitates for the rise, spread, and possible cause of the new disease. Determining if the increase in patients is related to this new infectious disease. Calculating if it is targeting specific age groups.

The data set consists of 65 patients that have the infectious disease with ages ranging from 35 years of age to 81 years of age for NCLEX Memorial Hospital. Remember this assignment will be completed over the duration of the course.

**Classify the variables in your data set.**

* Which variables are quantitative/qualitative?
  + Quantitative variable would be the age variable
  + Qualitative would be the patient numbers of the sample
* Which variables are discrete/continuous?
  + Patient numbers are discrete
  + Patient age is continuous
* Describe the level of measurement for each variable included in your data set.
  + There are 65 patients being used in this sample training from ages 35 to 81.
* Discuss the importance of the Measures of Center and the Measures of Variation.
  + the measure of center summarizes in variable a representation of all the data from its center. Variation uses this center to find normal distributions as a way to expresses the standard deviation as a percentage of the mean.
* What are the measures of center and why are they important?
  + Mean and median both try to measure the "central tendency" in a data set. The goal of each is to get an idea of a "typical" value in the data set. The mean is commonly used, but sometimes the median is typically preferred. It is resistant to the presence of extreme values
* What are the measures of variation and why are they important?
  + The measures are the range, standard deviation, and variance the range of a set of data values is the difference between the maximum data value and the minimum data value. The standard deviation of a set of sample values, denoted by s, is a measure of how much data values deviate away from the mean. basically, a transformation or velocity from the mean. The variance is a measure of how spread out the set of data is from the average (mean, median).

Calculate the measures of center and measures of variation. Interpret your results in context of the selected topic. Based off the age being they all have the infection and the numbers to identify them are not quantitative.

~~Mean 63~~

~~Median 64~~

~~Mode 69~~

~~Midrange 58~~

~~Range 46~~

~~Variance 86~~

~~Standard Deviation 9.26~~

**Conclusion**

Recap your ideas by summarizing the information presented.

It seems that the age average age range of patients at a median of 64 years old and the mean of 63 years of age have the average infection rate meaning this age range could more susceptible to the infection.it seems the standard deviation is spread apart at 9.26 which means based on the sample of patients that age range of infected patients varies more than those of the median or mean age range of patients at 63-64.

04/13/2020 Re calculated data statistics – Shaun Pritchard

|  |  |  |  |
| --- | --- | --- | --- |
| ***Age range of patients with infections disease sample (Sample Size 65)*** | | | |
|
| Stat | Value | Rounded | Population |
| Mid-Range | 58 |  |  |
| Mean | 62.53846154 | 63 |  |
| Standard Error | 1.148175167 | 1 |  |
| Median | 64 | 64 |  |
| Mode | 69 | 69 |  |
| Standard Deviation | 9.256884133 | 9 |  |
| Sample Variance | 85.68990385 | 86 | 86.37 |
| Kurtosis | -0.106939217 | 0 |  |
| Skewness | -0.600603229 | -1 |  |
| Range | 46 | 46 |  |
| Minimum | 35 | 35 |  |
| Maximum | 81 | 81 |  |
| Sum | 4065 | 4065 |  |
| Count | 65 | 65 |  |
| Confidence Level(95.0%) | 2.293743579 | 2 |  |

# Course Project – Phase 2

**Discuss the importance of constructing confidence intervals for the population mean.**

Essentially constructing confidence intervals give us the best point estimate or approximation of a population mean. With this, we can determine whether inference we are looking for is within the range of viability for being true or close to approximation as they pertain to said population means. We can find a range or an interval with a percentage that determines the deviations or approximation of the deviation from this true population.

Confidence intervals help us determine the margin of error which aid in approximation being split evenly into the trials of a normal distribution. The main point in which confidence intervals facilitate is determined the overall viability of statistics for greater proportion using just a small proportion sample from it. This save in computation when dealing with large populations in the real world.

For instance, if we want to determine how many drivers in America use their blinkers. It would make no sense to survey every driver in America 300+ million people to collect data on only to compute the probability in determining the event. We can take a sample from the population of Americans and survey this smaller amount to find a range bound by limits in which we could approximate how many drivers use blinkers.

The only catch is that the approximations for the distribution only apply to the sample population we tested. By doing so we can be confident that the approximation is within the upper and lower bounds of the test we run. It is quite an ingenious way to calculate and determine viability.

**What are confidence intervals?**

Confidence intervals are the intervals or limits in between which a specific chance that specific intervals contain the true mean. a range of values for a parameter that has an associated confidence level.

**What is a point estimate?**

It is a best guess or estimate of some value or parameter. The sample standard deviation, the sample mean, and sample variance are all point estimates. In the case of a confidence interval, it is the sample mean X̄

**What is the best point estimate for the population mean? Explain.**

The sample mean X̄ is the best point estimate for the population mean µ. We can calculate p the point estimate by driving the mean from the sample of the population to best approximate the true value of µ. To find the point estimate we need to calculate the point estimate +- the margin of error using the midpoint formula. we would divide the upper and lower limit values of error after adding them by 2 to get the midpoint or point estimate value **p**

**Why do we need confidence intervals?**

Confidence intervals provide us with an upper and lower limit around our sample mean, and with this interval, we can then be confident to some degree, percentage, and approximation that we have captured the true population mean.

**95% Confidence Interval Experiment**

**Find the best point estimate of the population mean.**

The best point of estimate isX̄ = 62.53

**Construct a 95% confidence interval for the population mean. Assume that your data is normally distributed and σ, the population standard deviation, is unknown.**

In this case we will have to use the T statistic because sigma is unknown.

**Confidence interval formula:** X̄ **+-** t\* α/2 \*  s / √n

*(confidence value)* CI = 95%

*(Sample size)* n = 65

*(Sample Mean)* X̄ = 62.53

*(Compute alpha)* (α): 1 - (confidence level / 100) = 1 - 0.95 = 0.05

*(critical probability)* (p\*) = 1 – α/2 = 1 - 0.05 / 2 = 0.975

*(degrees of Freedom)* DF = (n-1) = 64

*(Sample* σ *)* s = 9.256884133

*(Standard Error)* SEx = s / sqrt( n ) = 9.256884133 / √65 = 1.148561642

*(critical t statistic value)* (t\*) = 1 – α / 2 \* P(DF) = invt(0.025,65) = -1.997729633

*(Margin of Error)* (ME) = Critical Value \* Standard Error = 2.29374

**Please show your work for the construction of this confidence interval and be sure to use the Equation Editor to format your equations.**

Confidence T level formula:

X̄±t\* α/2 \*  s / √n = 62.53 ± 2 \* 9.256884133 / √65 = **2.29**

Upper **=** X̄+t\* α/2 \*  s / √n = 62.53 + 2 \* 9.256884133 / √65 = 62.53 + 2.29 = **64.82**

Lower = X̄ - t\* α/2 \*  s / √n = 62.53 - 2 \* 9.256884133 / √65 = 62.53 - 2.29 = **60.23**

**Write a statement that correctly interprets the confidence interval in context of your selected topic.**

If we continue to repeat the experiment with the same sample size, we will get the same confidence level whereas the sample mean will fall into a range 95% of the time between the range 60.23 < mu < 64.82. With 95% confidence we can be sure the population mean will occur in this range.

**99% Confidence Interval Experiment**

**Find the best point estimate of the population mean.**

The best point of estimate isX̄ = 62.53

**Construct a 99% confidence interval for the population mean. Assume that your data is normally distributed and σ, the population standard deviation, is unknown.**

No population mean **μ**  or standard deviation **σ** is known so we must conduct a 2 tailed test using T critical values to find confidence values for the experiment.

**Confidence interval formula: X̄ +-** t\* α/2 \*  s / √n

*(confidence value)* CI = 99%

*(Sample size)* n = 65

*(Sample Mean)* X̄ = 62.53

*(Compute alpha)* (α): 1 - (confidence level / 100) = 1 - 0.99 = 0.01

*(critical probability)* (p\*) = 1 – α/2 = 1 - 0.01 / 2 = 0.005

*(degrees of Freedom)* DF = (n-1) = 64

*(Sample* σ *)* s = 9.256884133

*(Standard Error)* SEx = s / sqrt( n ) = 9.256884133 / √65 = 1.148561642

*(critical t statistic value)* (t\*) = p\* =1 – α / 2 \* P(DF) = invt (0.005,64) = -2.654854316

*(Margin of Error)* (ME) = Critical Value \* Standard Error = 3.04

**Please show your work for the construction of this confidence interval and be sure to use the Equation Editor to format your equations.**

X̄±t\* α/2 \* s / √n = 62.53 ± -2.655 \* 9.256884133 / √65 = **3.04**

Upper = X̄+t\* α/2 \* s / √n = 62.53 + -2.655 \* 9.256884133 / √65 = 62.53 + 3.04 **= 65.578**

Lower = X̄-t\* α/2 \* s / √n = 62.53 - -2.655 \* 9.256884133 / √65 = 62.53 - 3.04 **= 59.482**

It is 99% Confident the population mean is within the range: 59.48176 < mean < 65.57824

**Write a statement that correctly interprets the confidence interval in context of your selected topic.**

We can conclude that with 99% Confidence the population mean **μ**  is within the range: 59.48176 < mean < 65.57824. This experiment is conclusive to be within the range if we continue to repeat for the sample distribution.

**Compare and contrast your findings for the 95% and 99% confidence interval.**

We can be confident with 95% approximation that the population mean will be within the range 95% of the time between the range 60.23 < mu < 64.82. We can be confident that the mean would fall within the range: 59.48176 < mean < 65.57824 99% giving a higher approximation a higher margin of error at 3.04 as opposed to 2.29 with a 95% confidence level.

**Did you notice any changes in your interval estimate? Explain.**

Yes, I noticed the margin of error was largest for the higher percentage of confidence at 99% as opposed to the 95% confidence level. A meniscal fraction of the deviation became smaller segment in the tails when the percentage value was raised to 99% making the normal distribution wider.

Also, the closer we get to 100% the closer we get to 0 the CI coverage is much wider in relation to the 95% ME upper and lower bounds. It seems 95% get us closer to a smaller margin of error which a is more definitive approximation range.

**What conclusion(s) can be drawn about your interval estimates when the confidence level is increased? Explain.**

We can be confident with 95% approximation that the population mean will be within the range 95% of the time between the range 60.23 < mu < 64.82. We can be confident that the mean would fall within the range: 59.48176 < mean < 65.57824 99% giving a higher approximation a higher margin of error at 3.04 as opposed to 2.29 with a 95% confidence level. Based on the events of patients with infectious disease it would seem more probable that the average age of the patient who contracts the illness would fall within the age between 59 and 65 with only a .01% variance between the confidence level of 95% being patients from ages 60 to 65

# Course Project – Phase 3

1. Discuss the process for hypothesis testing.
   * **Discuss the 8 steps of hypothesis testing?**
     + **Step 1** - Identify the claim to be tested. This is also known as the alternative hypothesis. Where we specify the alternative indication that we want to use statistics to prove whether it is true or false.
     + **Step 2** - Establish symbolic form of notation based on the original claim when it is false. such as: EX: µ > than the H0, H0 = False.
     + **Step 3** - Identify what the null and alternative hypothesis are. We use the symbols below to describe them.

H0: µ = Some value

Ha: or H1: µ >,=,< some value

* + - **Step 4** - Determine the significance level. This is a statistical test that determines whether it will fall within the critical region if the null hypothesis is actually true. We denote this by using Alpha α = 0.5, or 0.01, etc...
    - **Step 5** - Identify the test statistic that we are going to use and determine the sampling distribution. This is the value we used to determine the decision about the null hypothesis. This is completed by converting the sample statistic to a score known as the Z-score test and/or the T-score test. We can determine which test we are going to use based on the availability of statistical information such as the population mean.
    - **Step 6** - Find the values of a test statistics and then we have to find either the P-value or the critical values we use a Point estimate to evaluate the population proportion. Then we can use the P-value and critical values for the test to identify the type of hypothesis test that we need to do. This helps us determine whether we are doing a two-tailed test or whether were testing for left or right-tailed test. The indications of these tests are based on whether we're trying to prove if something is equal which we would use a two-tailed test if it is greater than we would use a right-tailed test and if it is supposed to be less than the alternative claim of the null hypothesis we use a left tailed test. These tests determine the critical regions which we use as our basis to reject a hypothesis.
    - **Step 7** - Make our decision to either reject or fail to reject the alternative hypothesis. I should also note that there are 2 different errors we can calculate with our statistics. A type 1 error is the mistake of rejecting the null hypothesis when it is actually true. or a type II error where we fail to reject the null hypothesis when it is actually false.
    - **Step 8 -**  Restate the decision(claim) using simple non-technical to conclude a hypothesis test. We do this by claiming that there is ether no sufficient evidence to support the claim or that there is sufficient evidence to support the claim of the hypothesis test.
  + **When performing the 8 steps for hypothesis testing, which method do you prefer: P-Value method or Critical Value method? Why?**

Both tests have different approaches but essentially give you the same result. The p-value test has an advantage that you just need to compute the p-value to do the test. The P-value requires just one input to complete the computation. This is typically the most used approach for hypothesis testing in calculators and software.

The critical value test is the standard score that determines the area and one tell on the opposite side of the critical value or values from zero. It equals two corresponding significance level of alpha. The p-value is the probability of obtaining a test statistic for the sample from the population that assumes the null hypothesis is true.

The critical value text observes the test statistics and critical value. If the test is greater than or equal to the critical value, then we can conclude that the null hypothesis is rejected if the test statistic is less than the critical value then we conclude that the null hypothesis is not rejected.

When using a calculator (TI-84), it is much easier to determine the P-Value. For more accurate computation when proving a hypothesis on paper I think the critical value would be more preferable in determining more accurate statistics. We can obtain this using student Z or t-distribution tables. Using the Alpha level of significance and degrees of freedom (n-1) to pull the appropriate values.

1. **Perform the hypothesis test.**

The data set is selection #2 of patients with infectious disease below is my proof statistics of the hypothesis test.

* + - Original Claim: The average age of all patients admitted to the hospital with infectious diseases is less than 65 years of age.
    - Test the claim using α = 0.05 and assume your data is normally distributed and σ, the population standard deviation, is unknown.

H0:  µ = 65 years of Age = *(Null hypothesis / original claim)*

Ha:  µ < 65 years of Age = *(Alternative hypothesis of the original claim is less than population µ)*

**Values for calculations:**

σ, the population standard deviation, is **unknown** *(suggest t-test)*

**α** = 0.05 *(Level of significance)*

**n** = 65 *(Sample size)*

**s** = 8.92434 *(Sample σ)*

**X̄** = 62.53 *(Sample Mean)*

***DF*** = 64 (1 – n = 1 – 64)

Test Statistic: A sample size greater than 30 means we will use Student *t-test* statistics; also, no population standard deviation is known:

t=\frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}}\overset{H_0}{\sim}t_{n-1}

*t =* **X̄** *– µ / s/ √n*

*t = 62.53 – 65 / 8.92434 / √65 = -2.47 /.8620822517 = -2.151239699*

***t*** *= -2.151239699*

***p-value*** *= .0176 (The result is significant at p < .05)*

P(***t*** **α** | H0 is true)  = Cumulative distribution function of the distribution of the test statistic (***t***) under the null hypothesis.

*P(* ***t*** *<* **α** *) = P(.-2.15<0.05) = .0176 P is less than* **α** therefore we can reject the null hypothesis

**Steps to complete calculation with TI-84:**

STAT >> TEST >> T-Test >> Stats; enter in H0 = 65 , **α = 0.05, X̄ = 62.53, n = 65, <**µ0 🡪 calculate:

The ***t***-value statistic is within less of the critical region of **α, so the null is rejected at 5% level of significance.**

The average age of patients admitted to the hospital with this infectious disease is less than 65 years old based on sufficient evidence the null hypothesis can be rejected based on the basis of the p-value = .0176.

* + Based on your selected topic, answer the following:
    - **Write the null and alternative hypothesis symbolically and identify which hypothesis is the claim.**

H0:  µ = 65 if P-value is = **α,** fail torejectH0

Ha:  µ < 65 = true, if P-value is < **α,** rejectH0.

* + - **Is the test two-tailed, left-tailed, or right-tailed? Explain.**

This is a left tailed test because we were testing if Ha was less than <, H0:  µ = 65. Evidence shows that the P-value .0176 < **α = 0.005** and lies within the left critical region of an assumable normal standard distribution.

* + - **Which test statistic will you use for your hypothesis test; z-test or t-test? Explain.**

We used the T-test statistic because the sample size *n* was greater than > 30 and the population mean was unknown.

* + - **What is the value of the test-statistic? What is the P-value?**

***t*** *= -2.151239699*

***p-value*** *= .0176*

* + - **What is the critical value?**

Critical Value = 1.6690 = 1-SIDED TEST LEFT = 1.67

We can find the critical value by using the Student ***t*** distribution table with degrees of freedom df = n – 1 = 65 – 1 = 64, based on a left tailed test, with a level of significance of **α = 0.005** of the critical region.

* + - **What is your decision; reject the null or do not reject the null?**

We will reject the null hypothesis based on sufficient evidence.

* + - **Explain why you made your decision including the results for your p-value method or the critical value method.**

*P(* ***t*** *<* **α** *) = P(.0176<0.05) P is less than* **α** therefore we can reject the null hypothesis. I did not have to look up the t distribution values for this conclusion.

* + - **State the final conclusion in non-technical terms.**

We have sufficient evidence to support the claim that sample mean is less than the average age of patients 65 years old who were admitted to the hospital with this infectious disease.

# Course Project – Phase 4

**Conclusion:**

The initial determination of this proof was to determine which patients based on their age correlated or facilitated the rise of spread the possible cause of a new disease. From a dataset sample mean of 65 patients from ages ranging 35 to 81 years of age. All known to have contracted the disease was used to run statistics test and determine the average age range are patients who are more susceptible to this infectious disease. This test statistic was also used to determine what significant evidence the probability-based off the sample to suggest a facilitated determination facilitate the determination of the overall average age range within the population who are more susceptible to contracting this disease.

We gathered 3 types of statistical analysis & test:

* **Descriptive statistics**
* **Inferential confidence intervals statistic**
* **Inferential hypothesis testing probability statistics**

The **Descriptive statistics** variables for performing calculations and probability tests in *Figure: 1*, below basic . these statistics where calculated to give of the basic middle numbers, sample variants, and sample standard deviation. These statistics would be used to conduct inferences and probabilistic testing.

**Descriptive statistics** *Figure: 1,*

|  |  |  |  |
| --- | --- | --- | --- |
| ***Age range of patients with infections disease sample (Sample Size 65)*** | | | |
|
| Stat | Value | Rounded | Population |
| Mid-Range | 58 |  |  |
| Mean | 62.53846154 | 63 |  |
| Standard Error | 1.148175167 | 1 |  |
| Median | 64 | 64 |  |
| Mode | 69 | 69 |  |
| Standard Deviation | 9.256884133 | 9 |  |
| Sample Variance | 85.68990385 | 86 | 86.37 |
| Kurtosis | -0.106939217 | 0 |  |
| Skewness | -0.600603229 | -1 |  |
| Range | 46 | 46 |  |
| Minimum | 35 | 35 |  |
| Maximum | 81 | 81 |  |
| Sum | 4065 | 4065 |  |
| Count | 65 | 65 |  |
| Confidence Level(95.0%) | 2.293743579 | 2 |  |

**Inferential confidence intervals statistic**

We then use these variables to conduct confidence interval experiments to determine a margin of error and the upper and lower limits of the age ranges of patients more likely to contract this infectious disease. Assuming our data was normally distributed, and the population standard deviation was unknown. The tests we conducted were at 95% and 99% intervals.

**95% Confidence Interval**

Upper = X̄ + t\* α/2 \* s / √n = 62.53 + 2 \* 9.256884133 / √65 = 62.53 + 2.29 = 64.82

Lower = X̄ - t\* α/2 \* s / √n = 62.53 - 2 \* 9.256884133 / √65 = 62.53 - 2.29 = 60.23

**99% Confidence Interval**

Upper = X̄ + t\* α/2 \* s / √n = 62.53 + -2.655 \* 9.256884133 / √65 = 62.53 + 3.04 = 65.578

Lower = X̄ - t\* α/2 \* s / √n = 62.53 - -2.655 \* 9.256884133 / √65 = 62.53 - 3.04 = 59.482

The confidence tests were conducted to determine a certainty value of these limits based on a 95% confidence level and a 99% confidence level. Which would determine with comp the certainty of which the population mean would occur between these limits to determine the population range estimate.

**Evaluation:**

If we continue to repeat the experiments with the same sample size, we determined to get the same confidence level whereas the sample means will fall into the interval test.

At 95% confidence level we determined that the range 95% was between the range 60.23 < mu < 64.82. With 95% confidence that the population means will occur in this range.

At 99% confidence level we determined that the range 99% was between the range 59.48176 < mu < 65.57824 99% With 99% confidence that the population means will occur in this range.

**Determination:**

The determination noticed the margin of error was largest for the higher percentage of confidence at 99% as opposed to the 95% confidence level. A meniscal fraction of the deviation became smaller.

The lowest limit of both test being 59.482 and the upper limit being 65.578 would determine that The average patient who is susceptible to conducting this infectious disease would be more likely with certainty to be of the age range between 59 and 65 years old.

**Inferential hypothesis testing probability statistics.**

After determining this population range estimate. The variables calculated in each statistic where used to conduct a one variable probability hypothesis tests to determine if the average of all patients admitted with this infectious disease were less than the age of 65 years. Assuming a normal distribution and without knowing the population standard deviation.

**The claim established was as follows:**

H0:  µ = 65 years of Age = *(Null hypothesis / original claim)*

Ha:  µ < 65 years of Age = *(Alternative hypothesis of the original claim is less than population µ)*

The claim was to determine with sufficient evidence if the claim were true or false and if the null hypothesis would be rejected or fail to be rejected. It was tested using α = 0.05 and assume the data was normally distributed and σ, the population standard deviation, was unknown. Due to the n sample size is larger than 30 and not having the σ, the population standard deviation. This statistic test was conducted, and new variables were found to determine the interval percentage of the critical value and p-value with a t-distribution test.

**The new variable values of the test statistic were calculated as follows:**

t = -2.151239699

p-value = .0176

The Critical Value = 1.6690 was determined by using the T distribution table. A left tailed test was conducted because we were testing if Ha was less than <, H0: µ = 65. Evidence shows that the P-value .0176 < α = 0.005 and lies within the left critical region of an assumable normal standard distribution.

It was determined that the null hypothesis would be rejected based on sufficient evidence. Also, that sufficient evidence supported the claim that the sample mean was less < than the average age of patients 65 years old you are admitted into the hospital with this infectious disease

So, I can conclude with sufficient evidence that those patients slightly younger than the age of 65 up to the age of 65 are more vulnerable and susceptible to contracting this infectious disease. I would also conclude that patients that are of the age of 59 to 65 are more at risk with 95% certainty they will or could contract this infectious disease.

# References

Kahn Academy . (2020). *Confidence intervals*. Retrieved from https://www.khanacademy.org/: https://www.khanacademy.org/math/ap-statistics/estimating-confidence-ap/one-sample-t-interval-mean/v/calculating-a-one-sample-t-interval-for-a-mean

Mario F. Triola. (2018). Elementary Statistics. In M. F. Triola, *Elementary Statistics.* Pearson.