DS 6306 – Live Session 1

Answers to asynch questions

Christopher “Todd” Garner

Spring 2023

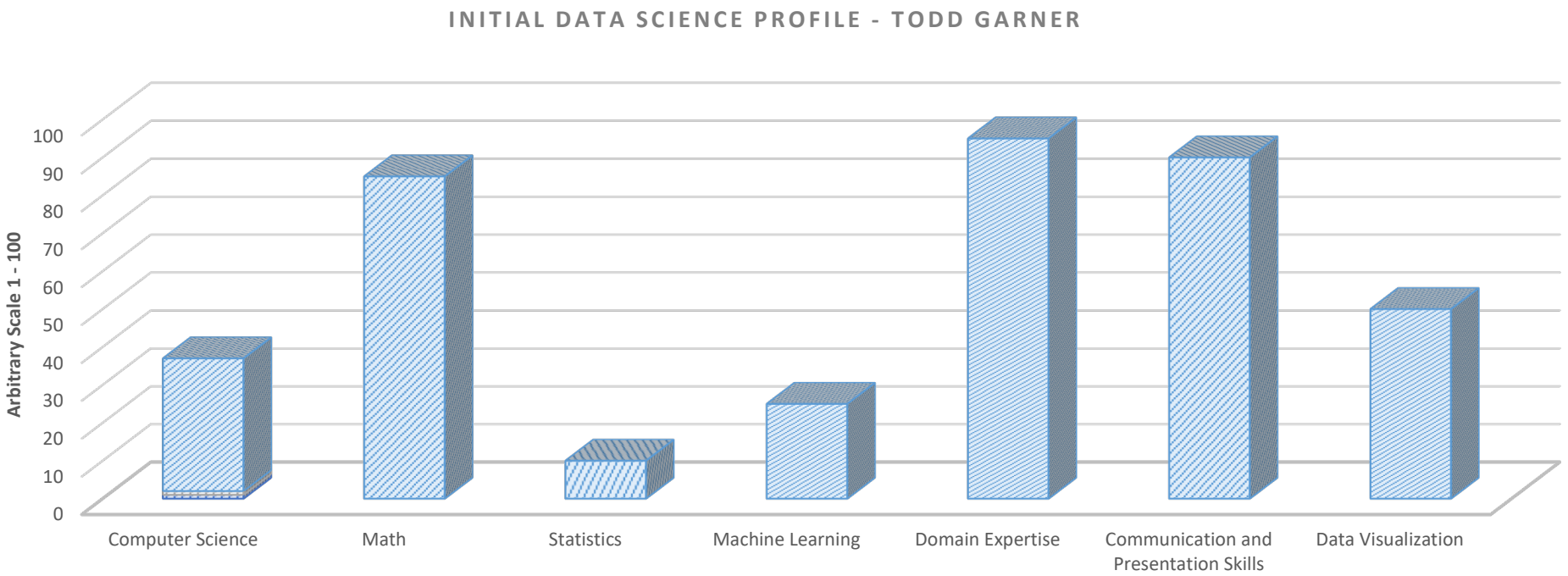
Question 1 - Data Science Profile - mine

- Question 1: Data Science Profile: Make a bar plot for your data science profile.
- Response: A very stimulating and subjective construction of major themes. Utilizing the major “topics” outlined in chapter 1 of *“Doing Data Science”,* the bar chart that follows was constructed.
- The major topics:
  - *Computer Science*
  - *Math*
  - *Statistics*
  - *Machine Learning*
  - *Domain Expertise*
  - *Communication and Presentation Skills*
  - *Data Visualization*

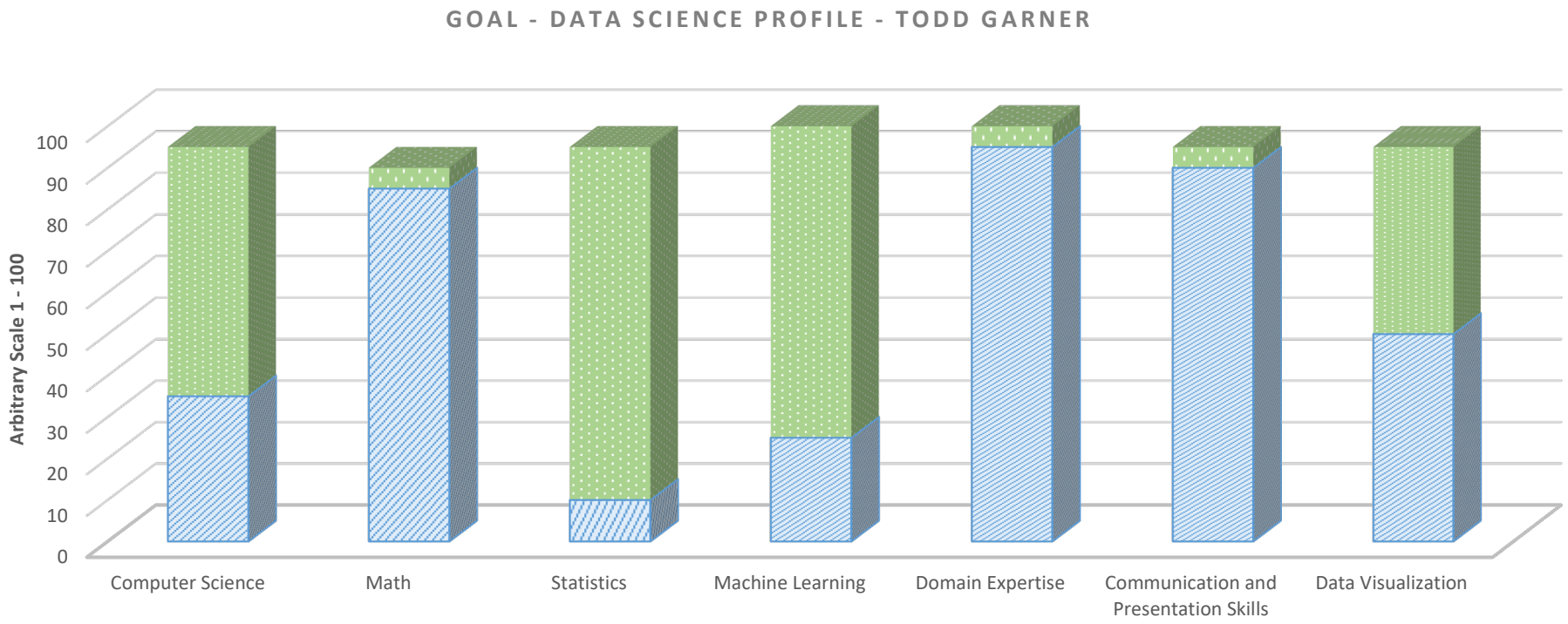
Data Science Profile – ‘cont

- While the major topics were variables capable of being displaced by other features, for simplicity, I chose to stay with the base topics shown in the book.
  - All the major topics were relatively straight forward, although still highly subjective. The outlier is Domain Expertise. Since I am constructing my profile, I’m going to define where I intend to focus throughout my Data Science journey, the energy sector.
  - I’m an oil and gas professional and I intend to take a deep dive into that industry. The use cases are endless and important. From selecting well locations (where to drill), model the appropriate height in the pay zone interval to land the lateral (HZ wells assumed), frac recipe (how much sand, proppant, etc.) and production profile (choke the production back, pump it out at max rate or other) to maximize the production of the well, over the life of the well/or, juice the NPV by pulling hard from the outset) to real-time equipment monitoring during operations to minimize downtime by predicting equipment failure, among many others. This is my passion to learn and perfect. This will be my domain.

Initial Data Science Profile



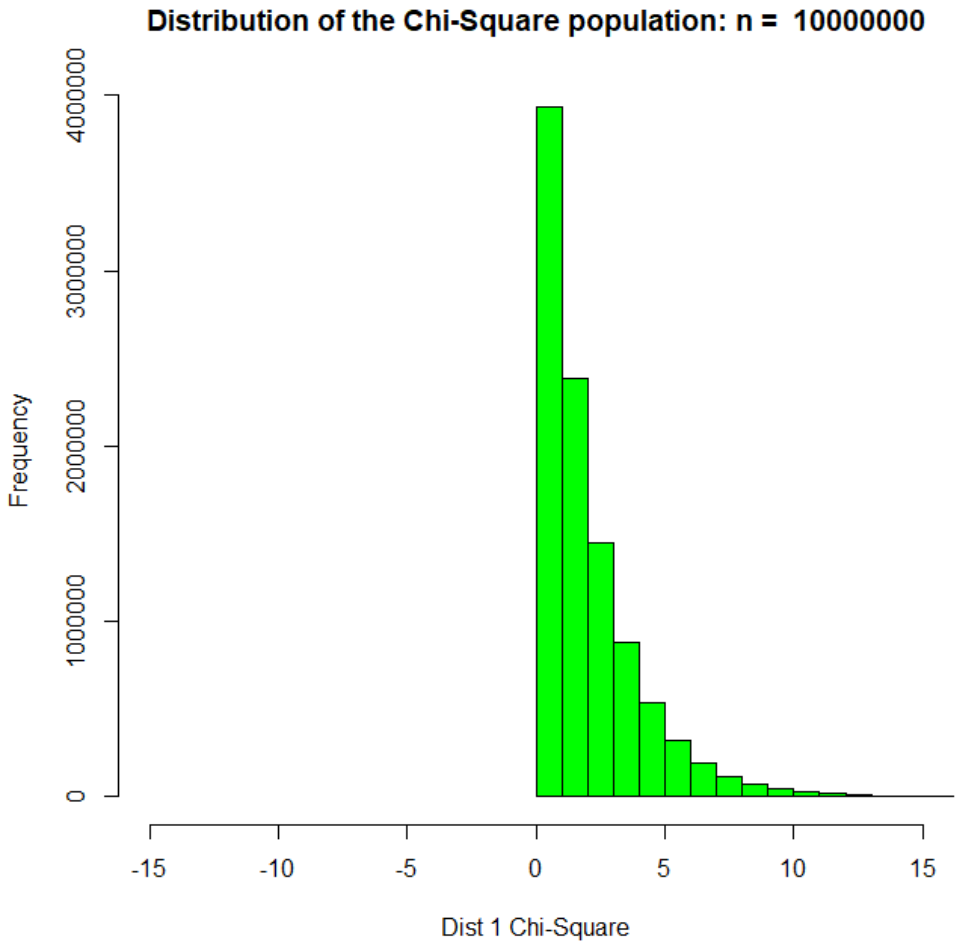
Goal – Data Science Profile



Question 2

- Part 1: Jumping into this piece of code was daunting and humbling. But I was able to prevail. I worked primarily in VS Code as a staging ground for a copy/paste operation into R. I did load RStudio and R Commander and spent some time gaining familiarity. I mostly watched them both become locked in an endless death loop but found ways to end those. I was able to strip away the unnecessary pieces of code and get it to a base set of code to achieve the desired results. On the next page, the histogram for a population of 10,000,000 from a Chi-Square distribution with 2 degrees of freedom is shown.
- UPDATE: After watching various videos on YouTube University, I get the distinction of R, knitr and Rmarkdown in RStudio. Genius, I look forward to utilizing these features.

Histogram - #2


Distribution of the Chi-Square population: n = 10000000
Frequency
0
1000000
2000000
3000000
4000000
-15
-10
-5
0
5
10
15
Dist 1 Chi-Square

Question #2, Part 3

- "The mean of the population is, 1.99878688861673 The standard deviation of the population is, 1.99931840113821“
- "The difference between the mean x and the standard deviation y is, -0.0005315125214806 48"

The Standard Deviation is slightly larger than the mean.

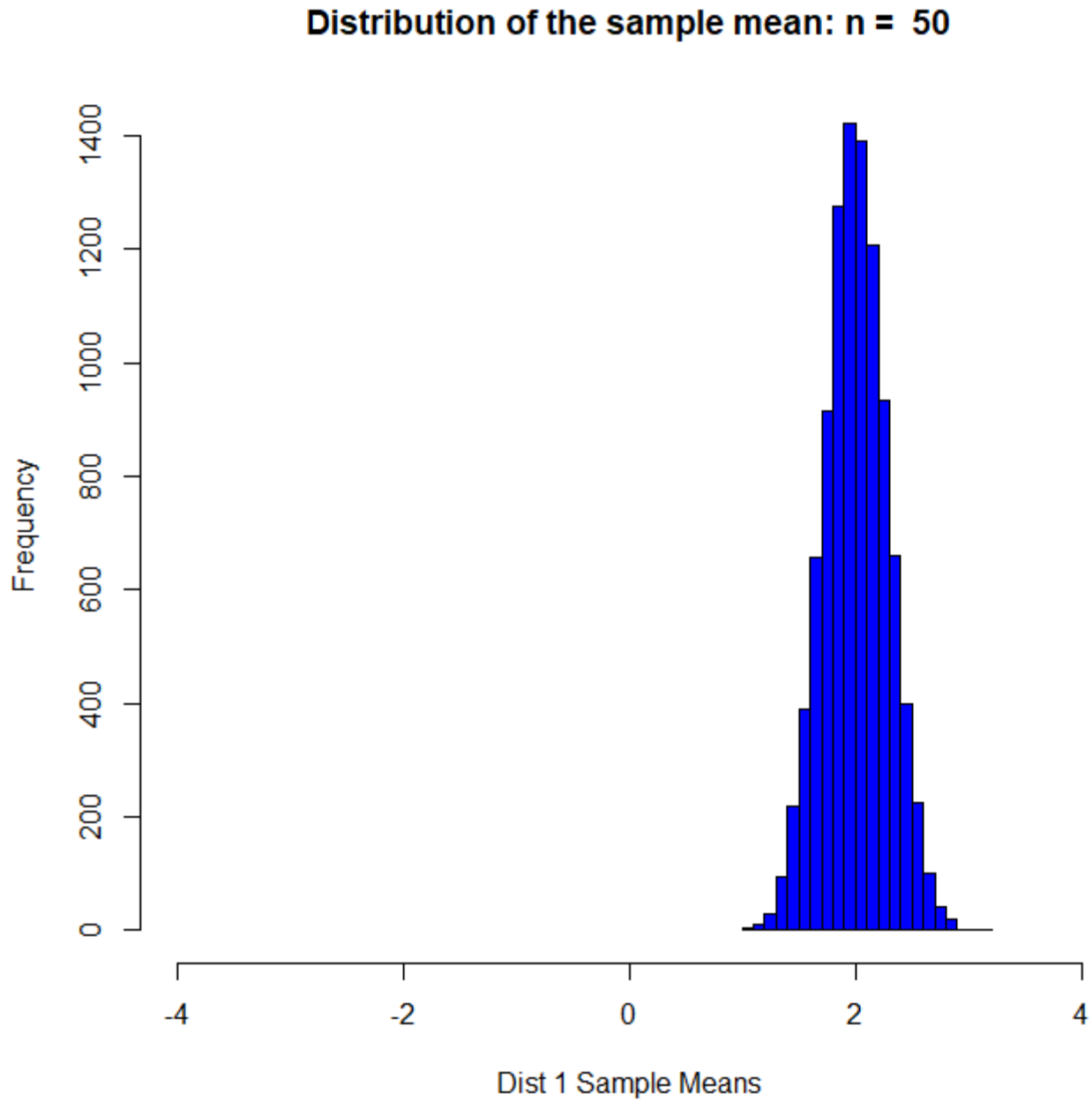
Question #2, Part 4

- Part 4 is the question: "According to CLT, what should be the approximate distribution of sample means of size 50 from this right skewed population?
- ANSWER: It should be normally distributed.
- What should be the mean and standard error of the mean (standard deviation of the distribution of sample means)?

ANSWER: Because the sample size is greater than 30, the mean of the sample size 50 (thus, normally distributed) should be the same as the mean of the population: $\mu_x = \mu_{sample}$ and the standard error should be smaller by a factor of $1/\sqrt{(50)}$

Question #2, Part 5

- Taking the average of the means from the sample size of 50, performed 10,000 times, is shown in the histogram on the next slide.
- This makes sense as the mean of the samples will remain the same (~2.0) as the mean of the population.
- However, the standard error of the distribution is smaller, by the factor (1/√(50)), which we can calculate as: 0.2790939 in R. By hand, sigma/ √(50) = .28274569. Reasonably close.
- Additionally, we can see that the means are normally distributed.


Distribution of the sample mean: n = 50
Frequency
0
200
400
600
800
1000
1200
1400
-4
-2
0
2
4
Dist 1 Sample Means

Question 2, Part 6

- Question: What is the mean and standard deviation of the 10,000 sample means.
- Answer (from R): Min. 1st Qu. Median Mean 3rd Qu. Max.
- 1.033 1.815 1.998 2.002 2.190 3.119
- Standard Deviation (from R): standard deviation of distribution 1
- [1] 0.2790939

Question 3 – T-test and six step hypothesis test for info below:

1 Hypothesis Test

The following are ages of 7 randomly chosen patrons seen leaving the Beach Comber in South Mission Beach at 7pm! We assume that the data come from a normal distribution and would like to test the claim that the mean age of the distribution of Comber patrons is different than 21. Conduct a 6 step hypothesis test to test this claim.

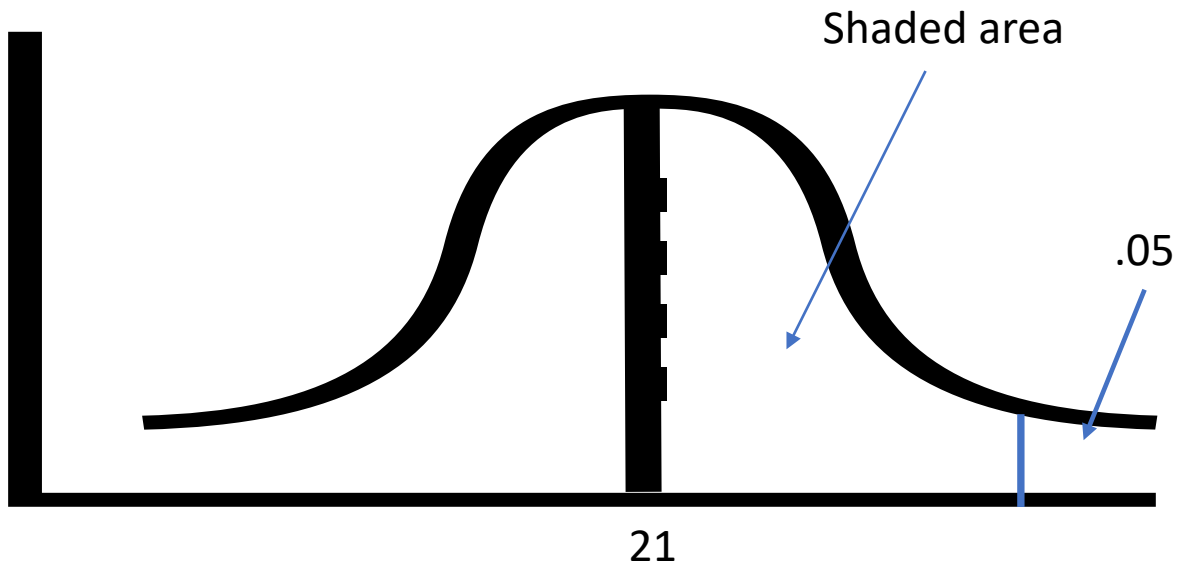
25, 19, 37, 29, 40, 28, 31

Question 3, T-Test – Step 1

- Step 1 – Formulate the Hypothesis:
- We want to test the Hypothesis that μ is not equal to 21
- Data – X_{bar} = 29.85714
  - μ_0 <= 21 H_o <= 21 H_o = 21
  - μ_a > 21 H_a > 21 H_a > 21

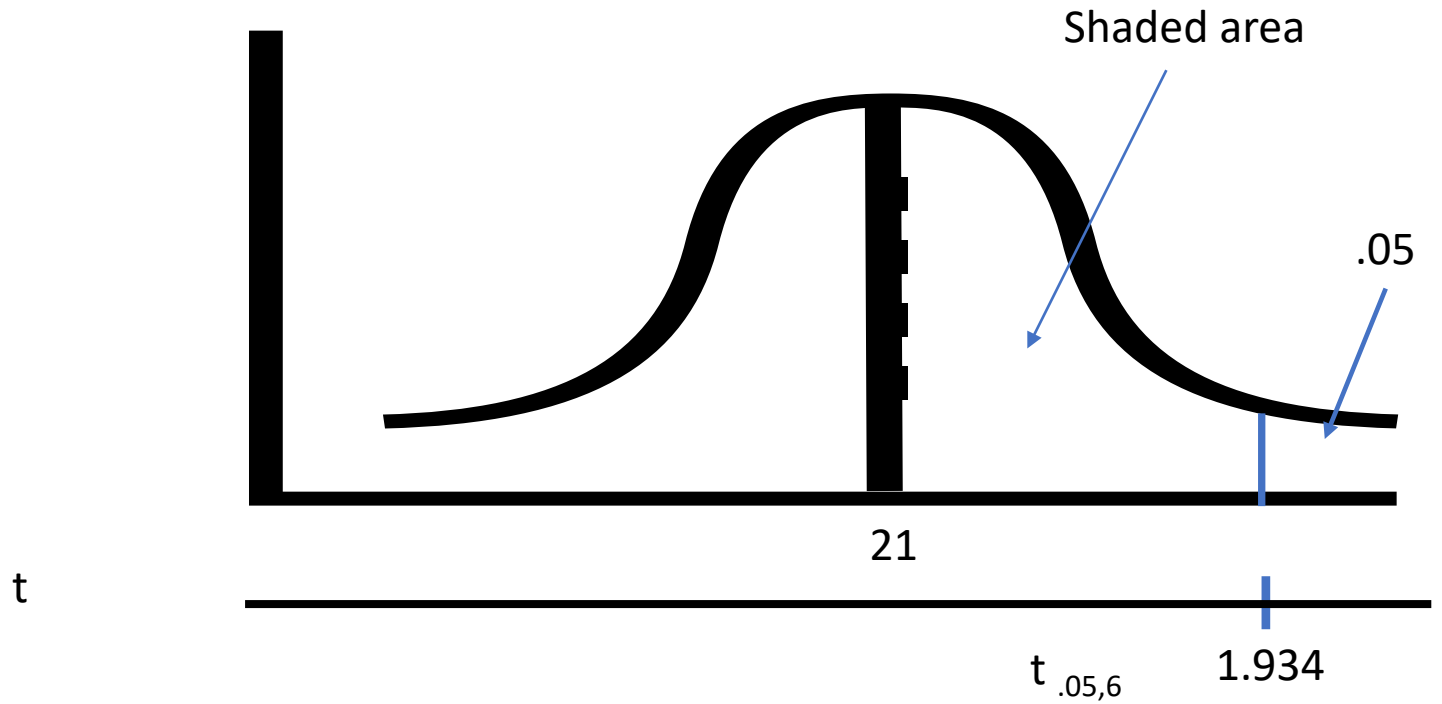
Question 3, T-Test – Step 2

- Draw and shade and find the critical value:
  - We don’t know that the values are normally distributed, and the sample size is only 7, less than 30.



Question 3, T-Test – Step 2

- To find the critical values we use the T-table:
  - T-value (Critical Value) from the T-table is:
    - 6 degrees of freedom
    - .05 one tail interval
  - T = 1.934

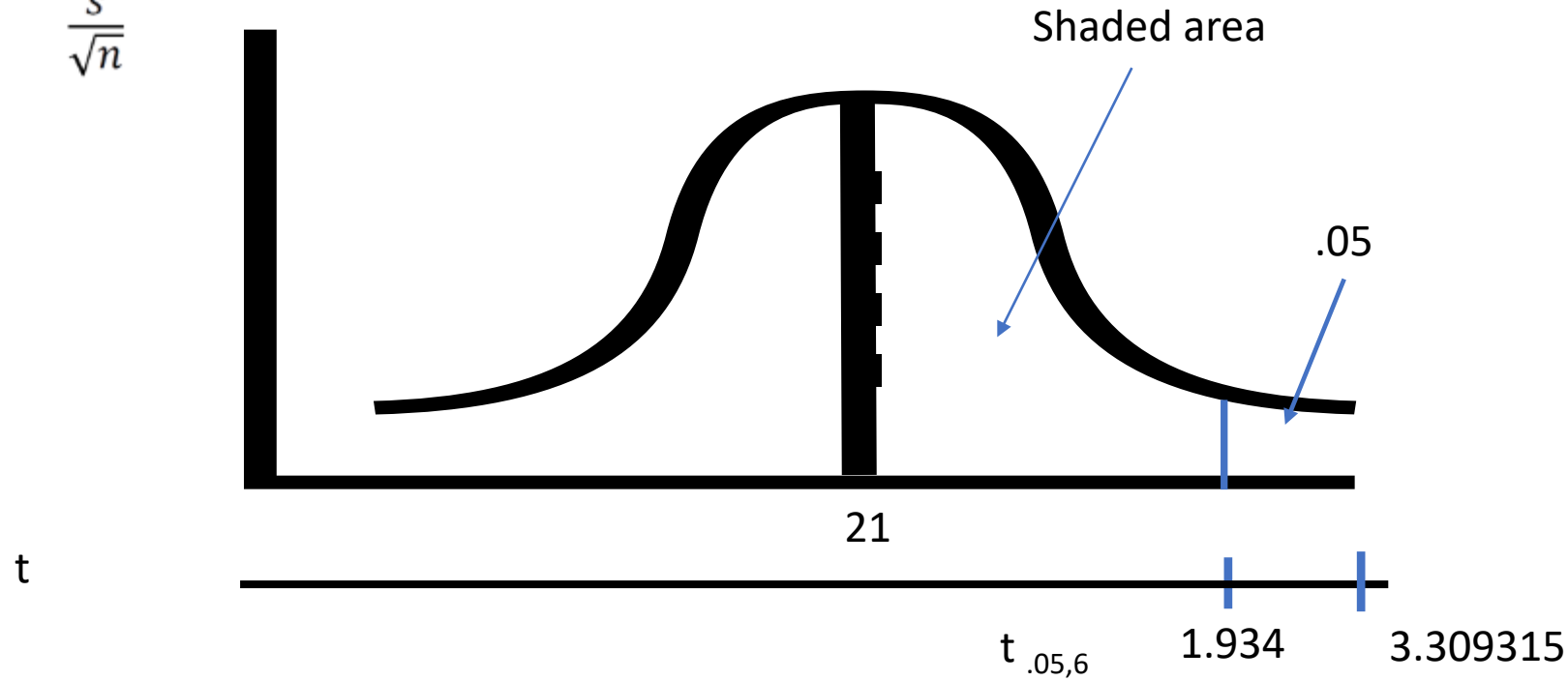


Question 3, T-Test – Step 3

- Step 3: Plug in the known values to calculate the t-score:
  - X_{bar} = 29.85714
  - μ_o = 21
  - s = 7.081162
  - n = 7

  - t = 3.309315

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$



Question 3, T-Test – Step 4

- Find the p-value: the probability of observing something as extreme or more extreme than what was observed, by random chance, under the assumption that the null hypothesis is true.
- p-value = .00810928

Question 3, T-Test – Step 5

- Alpha = .05
- p-value = . 00810928

- When the p-value is less than alpha, we must reject the null hypothesis H_o. In this case, p-value is less than alpha so we must reject the null hypothesis.

Question 3, T-Test – Step 6

- Conclusion: There is enough evidence, in fact strong evidence, that the true mean of the ages of the patrons of the Beach Comber are greater than 21. The p-value is 0. 00810928, which is significantly smaller than alpha (.05). The p-value is less than one sixth (1/6) of the value of the confidence interval. Therefore, we can confidently reject the null hypothesis that the average age of the patrons at the Beach Comber is 21.

Takeaways or Questions

Questions

- I grabbed a “normal distribution” from the libraries of “icons” in PPTx. Is there a better location for this type of art that can be shaded as needed?
- How about μ_{bar} (I did find a Greek letter generator online - https://www.greekalphabetletters.com/) but with the bar over the letter?

Takeaways or Questions

Takeaways/Comments:

- I’ve got a lot of learning to do. Which I find exciting! And daunting at the same time.
- I watched a very helpful YouTube video that was 45 minutes and got me from 0 to a working knowledge of RStudio and specifically R, knitr and Rmarkdown. https://www.youtube.com/watch?v=K418swtFnik&t=52s