Assignment 7: Pledge Individual Midterm

Instructions

- You must do this assignment by yourself. Do not discuss it with your classmates.
- Open notes. You may use the Internet to search, but you may not post questions and use ChatGPT.
- Produce your assignment as a qmd (or knitr) document rendered to pdf (knit to pdf).
- You will have to create your own qmd file for this midterm.
- Also submit your qmd file (it will not be graded but we want it for reference purposes).
- Show all the code as well as the results.
- 100 total points.
- Late submissions will be heavily penalized. Please give yourself ample time for knitting issues.
- For interpretation questions, refer to the TAs.
- Sign the honor pledge, the format of which is given below.

I, [insert name here], did not give or receive unauthorized aid on this exam.

Problem 1: [20 Points]

It is known that for $Z \sim N(0,1)$ and $V \sim \chi^2_{(n)}$ which are independent, then the transformation $T = \frac{Z}{\sqrt{V/n}}$ follows a student t distribution with n degrees of freedom. Select $n = 10$ to show that T follows student t distribution with $n = 10$ degrees of freedom. Use the Kolmogorov Sminov test to confirm this empirical distribution. Using ggplot, plot the empirical cumulative distribution function of T and theoretical t-distribution. Include the test statistic and p-value in your plot.

```r
set.seed(2023)
## Begin Solution



## End Solution
```

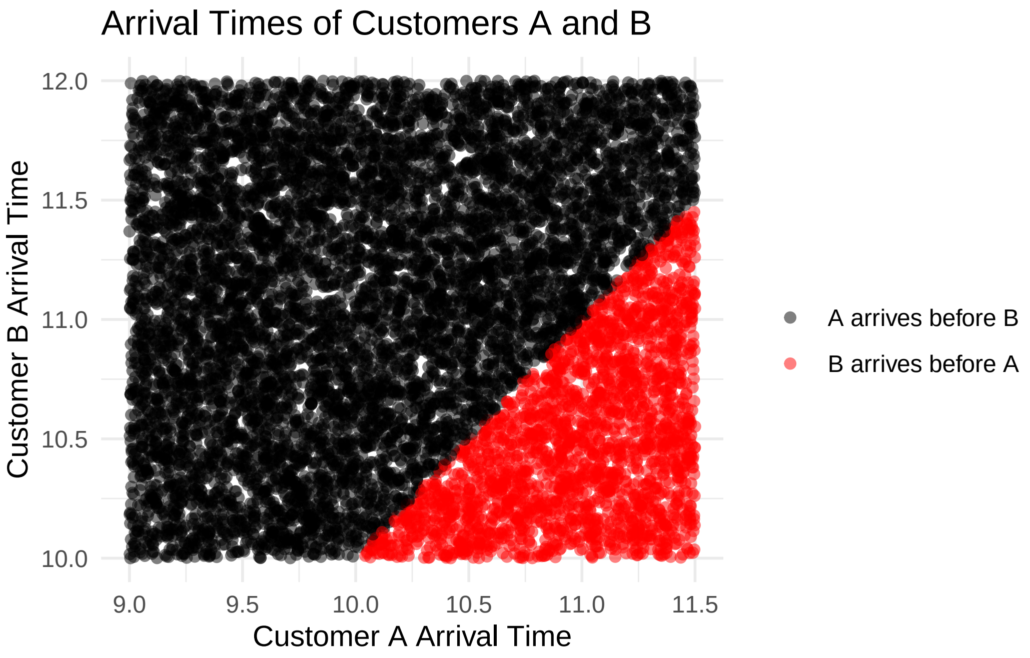
Problem 2: [20 Points]

Given that X_1 and X_2 represent the arrival times of customers A and B respectively to a store, and are independent and uniformly distributed, estimate the probability that customer B arrives before customer A using a simulation approach. Assume X_1 is distributed between time $t_0 = 9$ and $t_f = 11.5$ hours, representing the arrival time interval for customer A, and X_2 is distributed between time $k_0 = 10$ and $k_f = 12$ hours, representing the arrival time interval for customer B. Utilize a sample size of $n = 10000$ for the simulation.

Furthermore, compute the expected difference in arrival time, i.e., $E|B - A|$, and the variance of this difference.

Using ggplot, plot the paired arrival times of customers A and B, clearly highlighting the instances where Customer B arrives before Customer A. The outcome should be the following plot:

```r
set.seed(2023)
## Begin Solution



## End Solution
```



Problem 3: [30 Points]

Clarification: part 1 is worth 10 points and other parts are worth 5 points each

1. Write a program in R to simulate the students' performance and display the summary after each level, the following scenario is adapted from the CFA exam.

- 1. Assume there are 10 students and each student has an inherent proficiency, represented by a probability value that indicates their likelihood of correctly answering any given question. This probability ranges from 0.1 to 1.0. Generate these probabilities randomly for each student.(Hint: runif)
- 2. Create a function that simulates the grade of a student based on their probability. (Hint: rbinom)
- 3. Students undergo exams in three consecutive levels. For each level, the passing criteria are based on the relative performance of all participating students: Only the top 60% of students pass and move on to the next level.(Hint: quantile)
- 4. A student who doesn't pass can retake the exam once. After the retake, the top 60% of the failing students (based on the retake scores) will pass. If they fail again, they won't move to the next level. (Hints: logical vector, data.frame)
- 5. For each level, provide a summary: the amount of students who pass the current level and a table shows a). Then student's name (student_1,…,student_10) b). The student's inherent probability (show only two decimal places). c). Their grade.

d). Their grade on retake (if they took a retake, otherwise this should be NA). e). Whether they passed or failed. f). The number of times they retook the exam (0 or 1)

The summary format should similar to the follows:

```r
library(dplyr)
library(magrittr)
library(tidyr)
library(knitr)

set.seed(2023)
## Begin Solution



## End Solution
```

```
There are 8 students passed level 1

      student pass_fail probability grade retake_grade retake
1   student_1      pass        0.34    35           29      1
2   student_2      pass        0.43    41           44      1
3   student_3      pass        0.62    71           NA      0
4   student_4      pass        0.92    86           NA      0
5   student_5      fail        0.28    26           28      1
6   student_6      pass        0.91    88           NA      0
7   student_7      pass        0.95    99           NA      0
8   student_8      pass        0.69    68           NA      0
9   student_9      pass        0.67    66           70      1
10 student_10      fail        0.16    18           16      1

There are 7 students passed level 2

    student pass_fail probability grade retake_grade retake
1 student_1      fail        0.34    38           34      1
2 student_2      pass        0.43    50           44      1
3 student_3      pass        0.62    71           NA      0
4 student_4      pass        0.92    91           NA      0
5 student_6      pass        0.91    90           NA      0
6 student_7      pass        0.95    94           NA      0
7 student_8      pass        0.69    71           NA      0
```

```
8 student_9      pass        0.67    69            61      1

There are 6 students passed level 3

    student pass_fail probability grade retake_grade retake
1 student_2      fail        0.43    34            44      1
2 student_3      pass        0.62    72            NA      0
3 student_4      pass        0.92    90            NA      0
4 student_6      pass        0.91    89            NA      0
5 student_7      pass        0.95    94            NA      0
6 student_8      pass        0.69    69            61      1
7 student_9      pass        0.67    66            67      1
```

Using dplyr functions to solve part (2) to part (5).

2. Create a summary dataframe that consolidates the students' retake times and performance. Show a table with a format similar to the follows: (Hints: %>%, full_join, mutate, select, case_when, knitr::kable)

```
## Begin Solution



## End Solution
```

Table 1: Consolidated Student Performance

student	probability	retake	performance
student_1	0.34	2	fail level 2
student_2	0.43	3	fail level 3
student_3	0.62	0	pass the exam
student_4	0.92	0	pass the exam
student_5	0.28	1	fail level 1
student_6	0.91	0	pass the exam
student_7	0.95	0	pass the exam
student_8	0.69	1	pass the exam
student_9	0.67	3	pass the exam
student_10	0.16	1	fail level 1

3. Generate a summary dataframe that provides the total number of students and average probability for each performance status(show only two decimal places for avg_probability). Show a table with a format similar to the follows: (Hint: group_by,

summarise, knitr::kable)

```
## Begin Solution



## End Solution
```

Table 2: Performance Summary

performance	total_students	avg_probability
fail level 1	2	0.22
fail level 2	1	0.34
fail level 3	1	0.43
pass the exam	6	0.79

4. - 1. Create a new dataframe where you categorize students into different proficiency groups based on the probability values: High (probability >= 0.7), Medium (0.3 < probability < 0.7) and Low (probability <= 0.3).
   - 2. Modify the column names and select only relevant columns as the example table.
   - 3. Sort the dataframe based on the probability in ascending order.
   - 4. Show a table with a format similar to the follows:(Hint: %>%, mutate, rename, select, arrange, knitr::kable)

```
## Begin Solution



## End Solution
```

Table 3: Proficiency Level of Each Student

Name	Probability	Retakes	Performance	Proficiency
student_10	0.16	1	fail level 1	Low
student_5	0.28	1	fail level 1	Low
student_1	0.34	2	fail level 2	Medium
student_2	0.43	3	fail level 3	Medium
student_3	0.62	0	pass the exam	Medium

Name	Probability	Retakes	Performance	Proficiency
student_9	0.67	3	pass the exam	Medium
student_8	0.69	1	pass the exam	Medium
student_6	0.91	0	pass the exam	High
student_4	0.92	0	pass the exam	High
student_7	0.95	0	pass the exam	High

5.
   - 1. Add a new column, "Age" and assign age randomly from 18 to 30 for each student. Show the updated dataframe with table. (Hints: mutate, sample, knitr::kable).
   - 2. Select and display only the rows where students have passed the exam and are younger than 25. Show a table with a format similar to the follows: (Hints: knitr::kable, mutate, filter)

```r
set.seed(2023)
## Begin Solution



## End Solution
```

Table 4: Age of Each Student

Name	Age
student_1	26
student_2	21
student_3	24
student_4	18
student_5	19
student_6	30
student_7	24
student_8	28
student_9	19
student_10	28

Table 5: Students Who Passed the Exam and Are Younger Than 25

Name	Probability	Retakes	Performance	Proficiency	Age
student_3	0.62	0	pass the exam	Medium	22

Name	Probability	Retakes	Performance	Proficiency	Age
student_4	0.92	0	pass the exam	High	22
student_6	0.91	0	pass the exam	High	23
student_8	0.69	1	pass the exam	Medium	24

Problem 4: [30 Points]

- Clarification: each item is worth 7.5 points for STAT 405 and 5 points for STAT 605. Last item is required for STAT 605 and is extra credit for STAT 405. Just in case, keep in mind that if you are a 405 student that joined a group with a member that is a 605 student, you are NOT considered a 605 student for the purpose of this assignment (only for group related matters).
- To be a successful statistician, a very important concept you need to master is hypothesis testing. There are numerous resources on the Internet to consult if you need to refresh concepts. If you are not able in principle to obtain the same results as shown, it is very likely because you are missing some important theoretical aspect of hypothesis testing than to assume something is wrong with what is displayed in the plots.
- You will assume that you are sampling from a Normal distribution with known variance, so you will employ a z-test (not t-test).
- 605 students that extend the function to cope with unknown variance (hence using t-test) will have extra 5 credit points (not for 405, who have their chance to earn 5 extra points in the last item). For the t-test case assume the given standard deviation is the sample standard deviation (and add a parameter to the function to determine which version of test to use). If further things need to be assumed clearly state them. The plot will not be shown in this pdf. Be resourceful!
- The parameter and result values to use are:

```r
mu0 <- 4              ## Null hypothesis mean value
stdev <- 3            ## Known population standard deviation
signif.level <- 0.05  ## Test significance level
sample.mean <- 6.07   ## Mean of the random sample
n <- 10               ## Sample size
mu1 <- 6.2            ## Alternative hypotesis mean value to use
                      ## for error type 2 and power
```

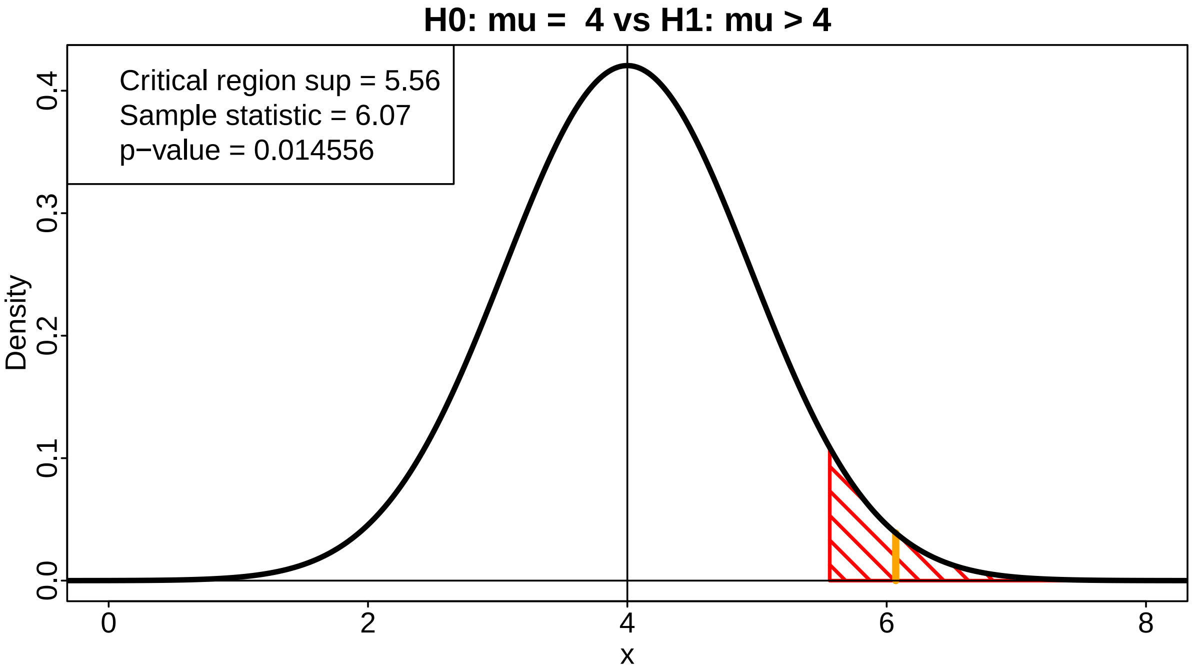
- What you will have to do is to replicate, using base R graphics, the following plots, that illustrate the key components of a test of hypothesis. Each plot needs to be produced by the same function, by changing the parameters of the function call. You must print each plot, not just the final plot.

- **STAT 405 students will have to produce only one-sided test plots.**
- **STAT 605 students will produce both one-sided and two-sided test plots.**
- You must produce a function with parameters that can be tweaked, such as:

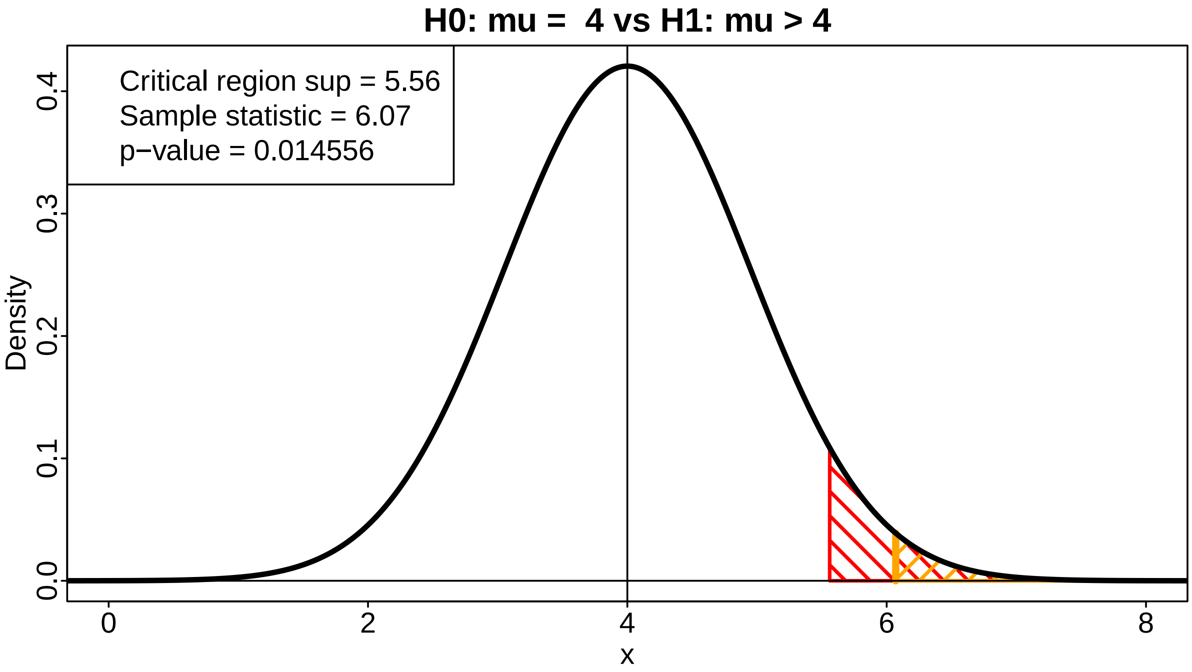
```
hyp.testing <- function(mu0, stdev, signif.level,
                        sample.mean, n,
                        show_crit, show_pvalue,
                        show_alt, mu1, show_beta, show_power,
                        two_sided) {
  ## 605 students that extend to t-test must add the parameter is_z_test
  ## so it is TRUE for z-test and FALSE for t-test

}
```

so you are able to produce the attached plots, using the function calls as shown

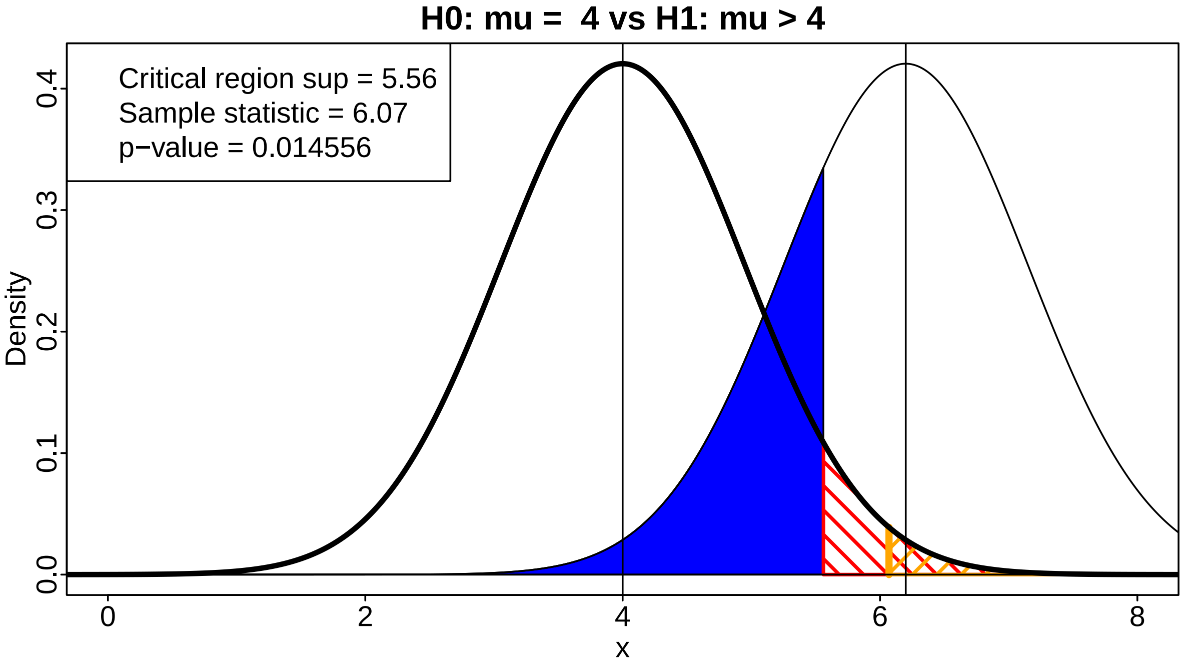
- 1. Density plot showing the critical region (red stripes. Hint: use `polygon`. It is also known as error of type 1 or α), and where the `sample.mean` (orange vertical line) is located:

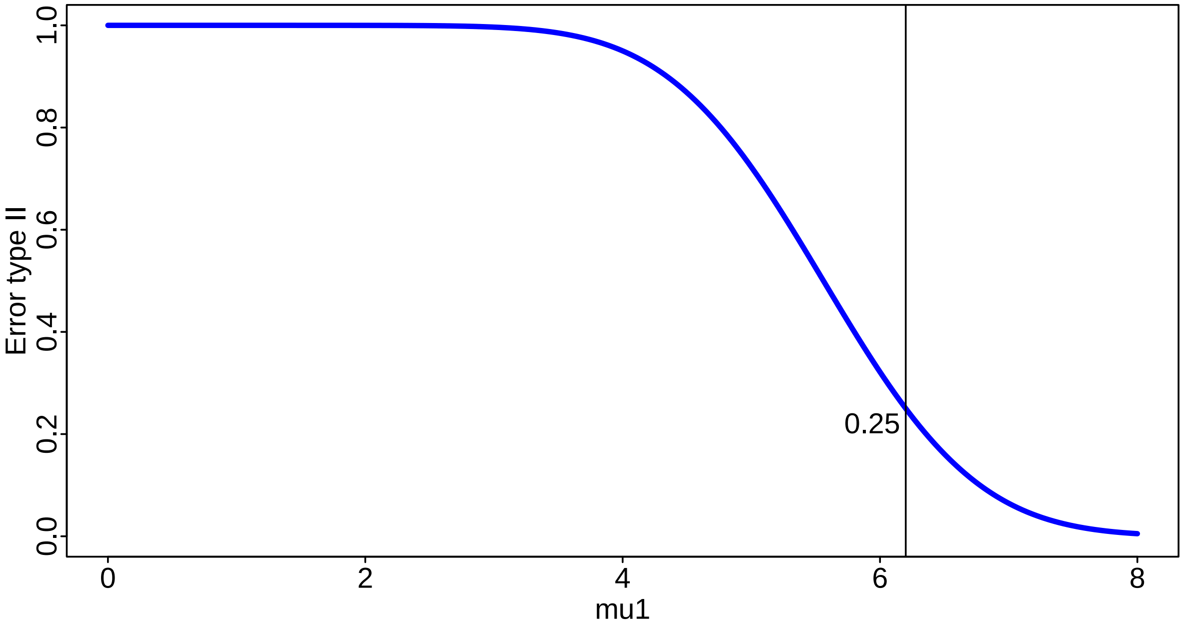


- 2. Superimpose the p-value probability region (orange stripes):

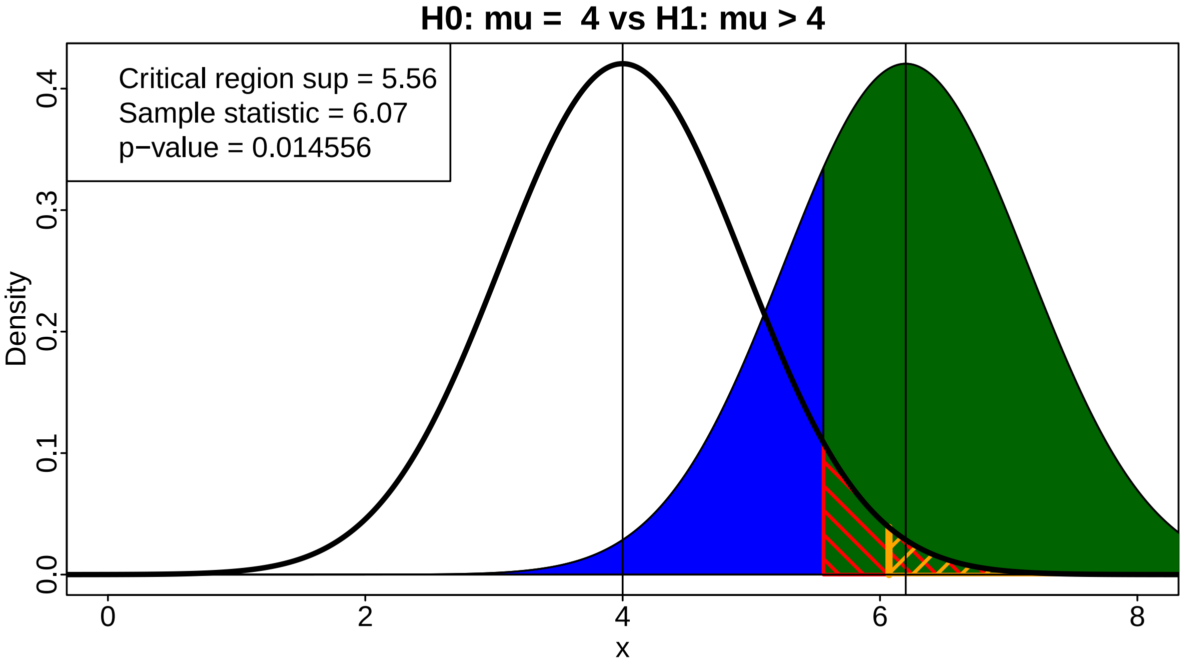


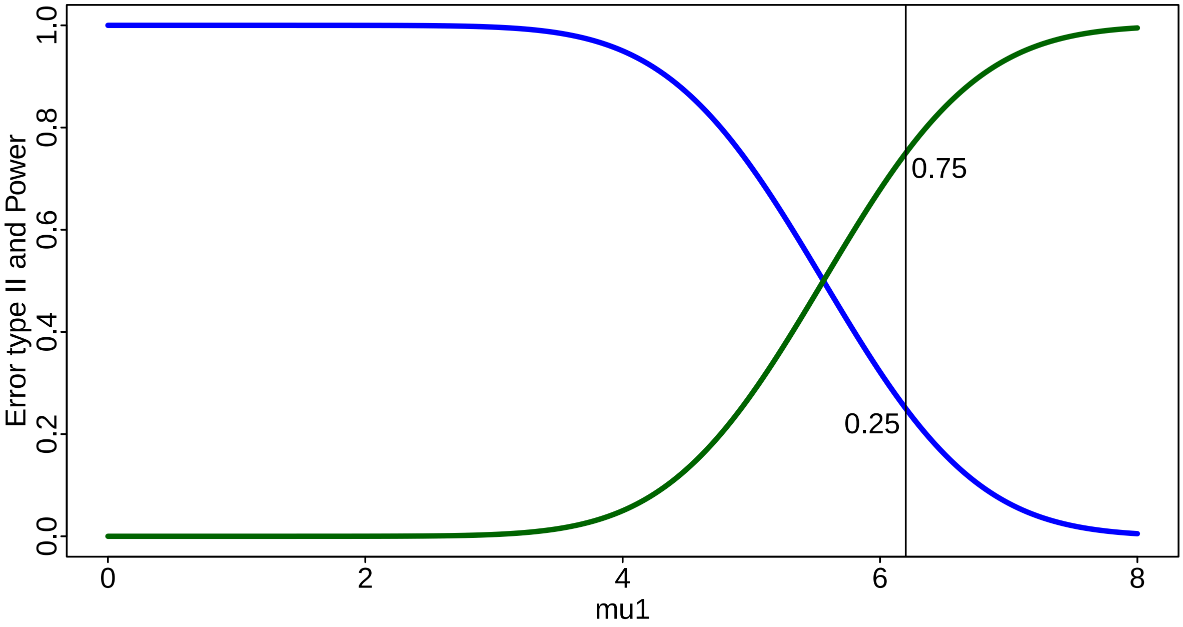
- 3. Add the error of type II (or β. Blue region and blue line):



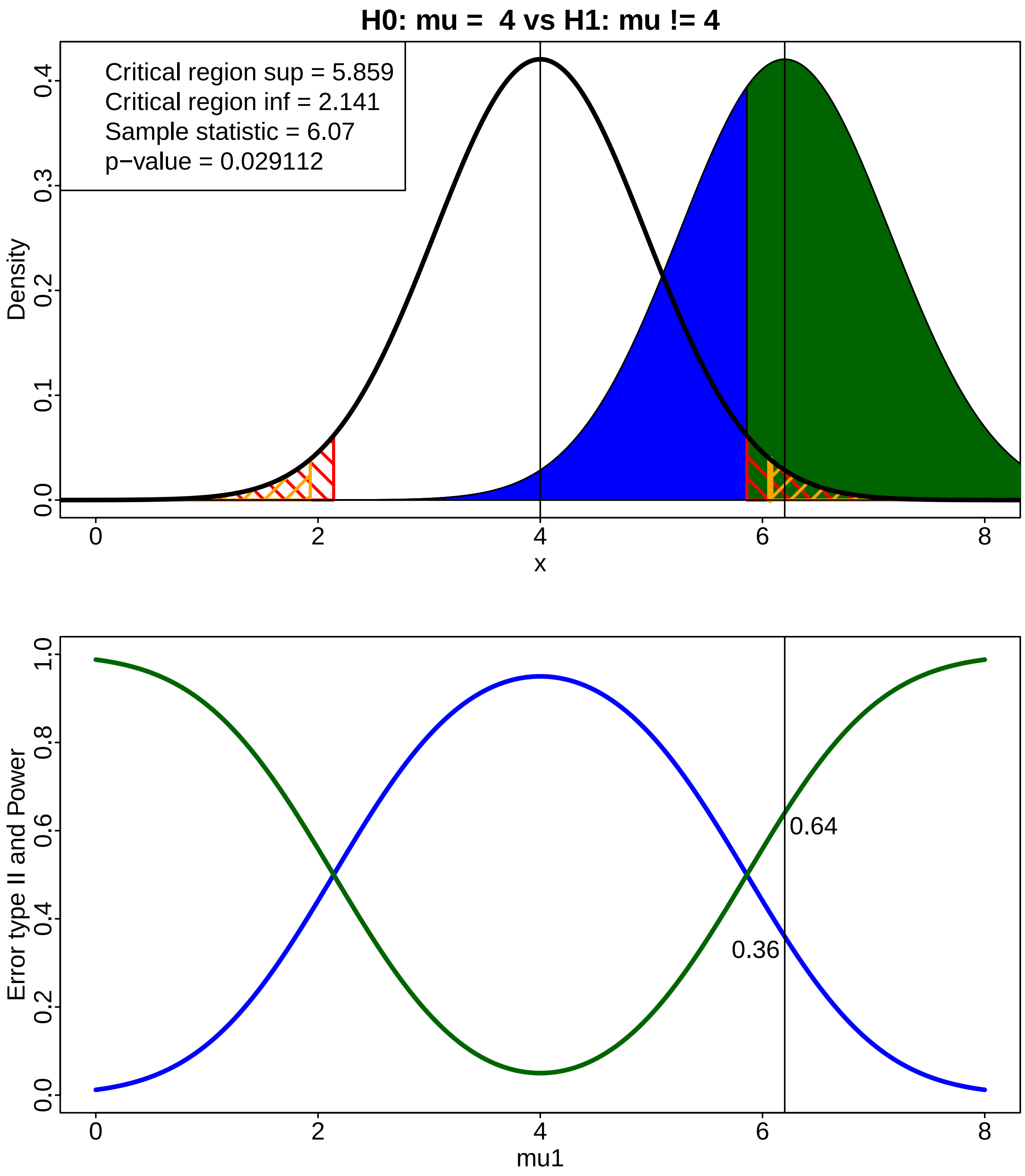


- 4. Add the power ($1 - \beta$. Dark green region and dark green line):





- 5. (STAT 605 required; STAT 405 for extra credit) Create the two-sided version.


H0: mu = 4 vs H1: mu != 4
Critical region sup = 5.859
Critical region inf = 2.141
Sample statistic = 6.07
p−value = 0.029112
Density
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
x
Error type II and Power
0.0
0.2
0.4
0.6
0.8
1.0
0.64
0.36
0
2
4
6
8
mu1