Worksheet 05 - Random Variables and Distribution Functions

##Shawn Kim

***Directions**: Please upload a PDF to Gradescope that includes both your written responses and corresponding R code inputs/outputs (if requested) for each problem. In code blocks, you can uncomment (delete #) the non-NOTE lines and fill in the blanks (FILL IN).*

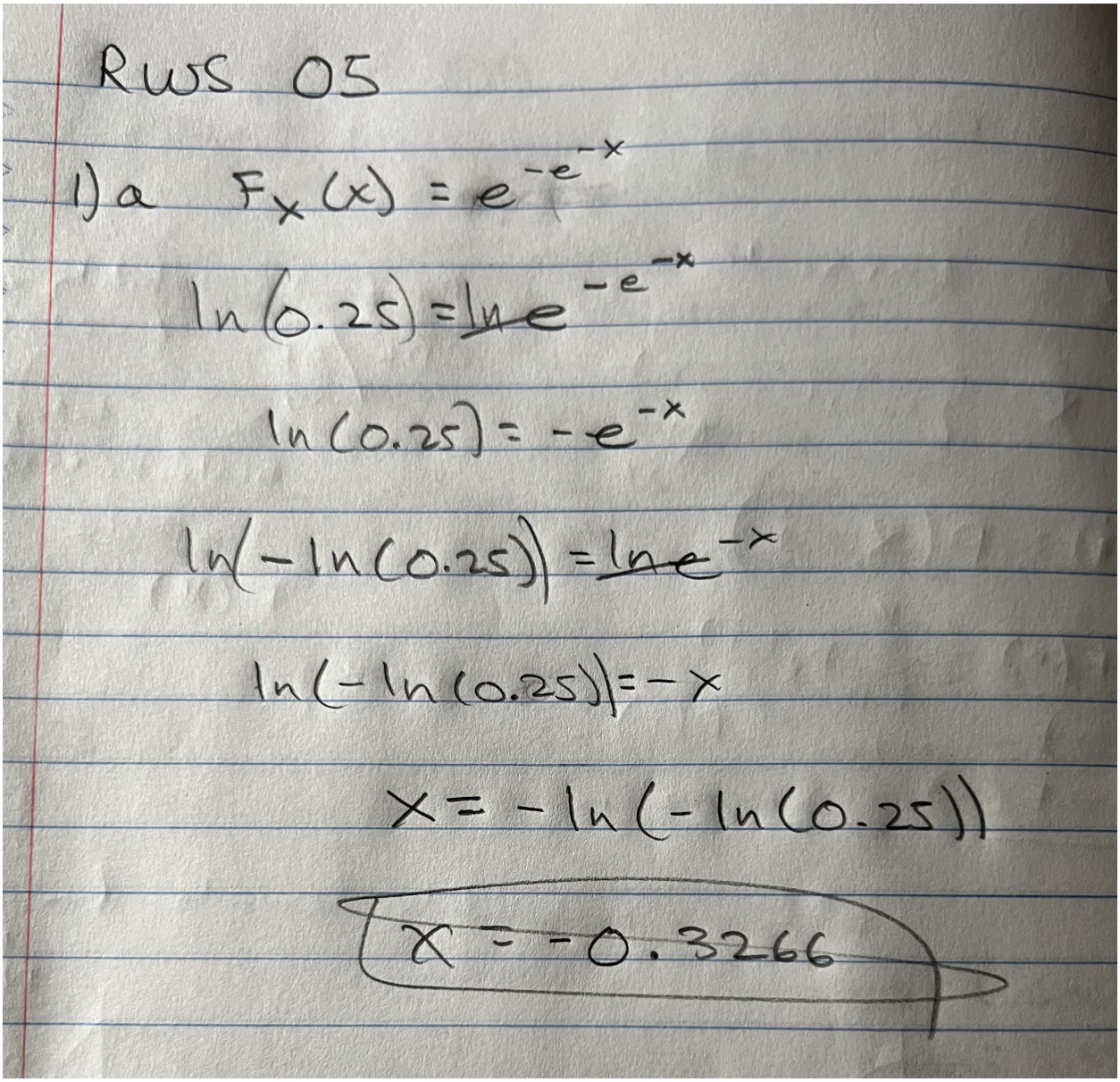
Problem 1. A Gumbel random variable X has the distribution function $F_X(x) = \exp(-e^{-x})$.

Problem 1 Part a) The 0.5 quantile of X occurs when $x = 0.3665$ since $F_X(0.365) = 0.5$. Additionally, the 0.75 quantile of X occurs when $x = 1.2459$ since $F_X(1.2459) = 0.75$. Put another way, the median and the third quartile of X are the values $x = 0.3665$ and $x = 1.2459$, respectively.

Confirm by-hand the calculation of the first quartile, $x = -0.3266$. Include all necessary steps in your calculation.

Hint: You will need to use a logarithm or two.

NOTE: Upload an image of your work.

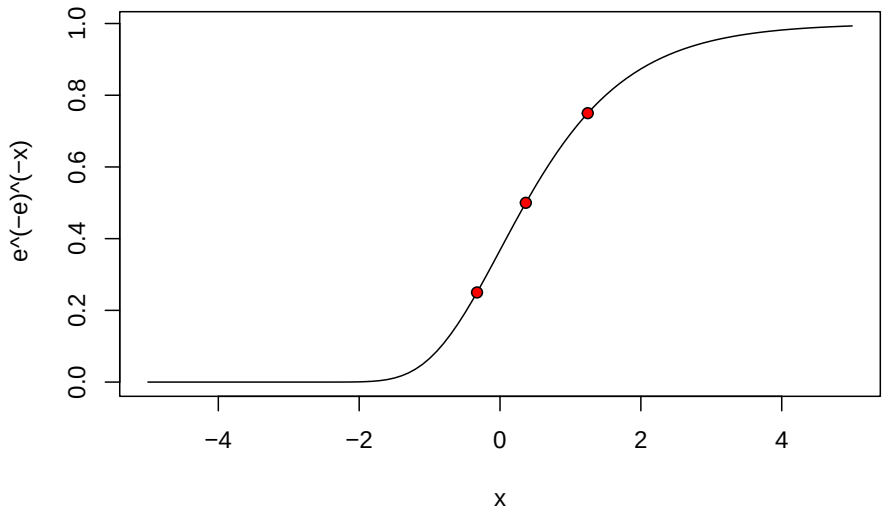


Problem 1 Part b) Using R, plot the graph of $F_X(x)$, indicating the first quartile, median, and third quartile as points on the plot. Label the axes appropriately. Below this plot, explain why F_X is a valid cumulative probability distribution function.

Hint: Remember to reference your notes so that you do not miss one of the properties of a valid cumulative probability distribution function.

```
# NOTE: The command curve(expression, from = NULL, to = NULL, xlab = xname,
# ylab = yname) draws a curve corresponding to a function defined by the
# expression over the interval [from, to]


# NOTE: For example, curve(x^2, from = -5, to = 4, xlab = 'x', ylab = 'y =
# x^2') plots the curve y = x^2 over the interval [-5, 4]
```

```r
curve(exp(-exp(-x)), from = -5, to = 5, xlab = "x", ylab = " e^(-e)^(-x)")

# NOTE: Refer back to your work in part a.  You may use the distribution
# function to solve for the quartiles in R (labeled 'points_xvals'), or you may
# simply use the calculations in part a. (It is good practice the both ways.)

# NOTE: Refer to Problem 2 Part d) in R Worksheet 01 for assistance with adding
# in the quartiles.

# NOTE: You may need to install the package evd.

library(evd)


x <- c(-0.3266, 0.3665, 1.2459)
y <- c(0.25, 0.5, 0.75)
points(x, y, pch = 21, bg = "red")
```



Space for Explanation:

F_X is a valid CDF becuase it is a non-decreasing function, approaches 0 as x approaches negative infinity, approaches 1 as x approaches positive infinity, and is right continuous

Problem 1 Part c) Consider the table of values for x and $F_X(x)$ for x equal to integers from -2 to 5, which has been generated in R.

```
##   x_val         F_X
## 1    -2 0.000617979
## 2    -1 0.065988036
## 3     0 0.367879441
## 4     1 0.692200628
## 5     2 0.873423018
## 6     3 0.951431993
## 7     4 0.981851073
## 8     5 0.993284702
```

Using the table above, determine the probability $P\{-1 < X \leq 4\}$ and $P\{X > 4\}$.

NOTE: Enter all necessary calculations as text or image.

$P\{-1 < X \leq 4\} = P\{X \leq 4\} - P\{X \leq -1\} = F_X(4) - F_X(-1)$

```
ys[7] - ys[2]
```

```
## [1] 0.915863
```

$P\{X > 4\} = 1 - P\{X \leq 4\} = 1 - F_X(4)$

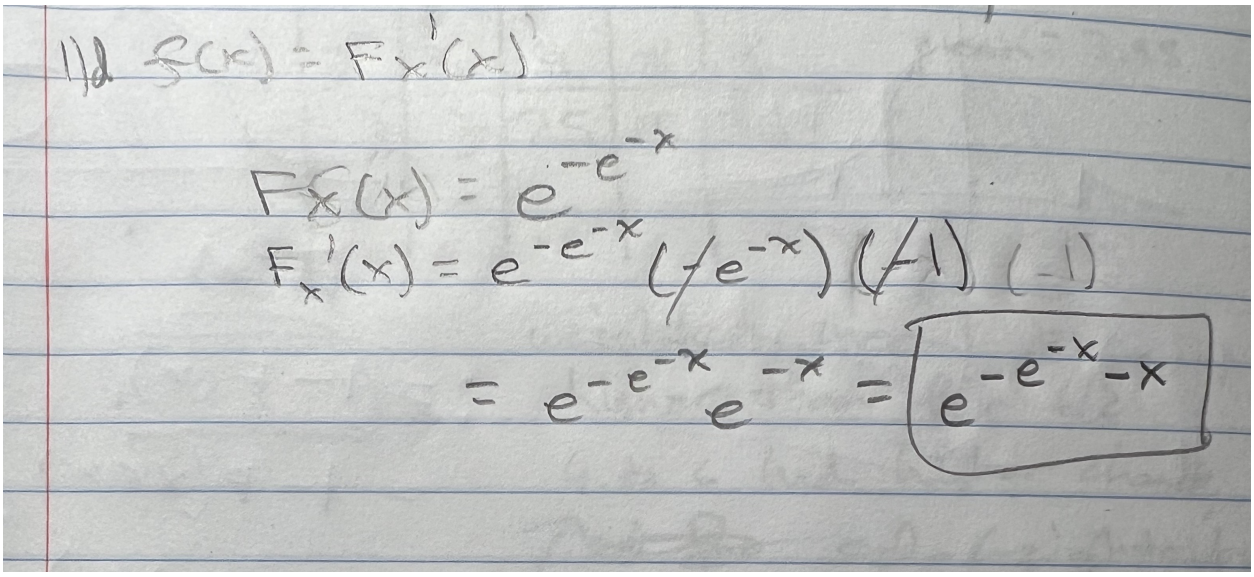
```
1 - ys[7]
```

```
## [1] 0.01814893
```

Problem 1 Part d) Determine the probability density function $f_X(x)$ for the cumulative distribution function $F_X(x) = \exp(-e^{-x})$.

Hint: The probability density function of a continuous random variable can be determined from the cumulative distribution function by differentiating using the Fundamental Theorem of Calculus, i.e., given $F_X(x)$, then $f_X(x) = \frac{d}{dx}F_X(x)$, as long as the derivative exists.

NOTE: Upload an image of your work.

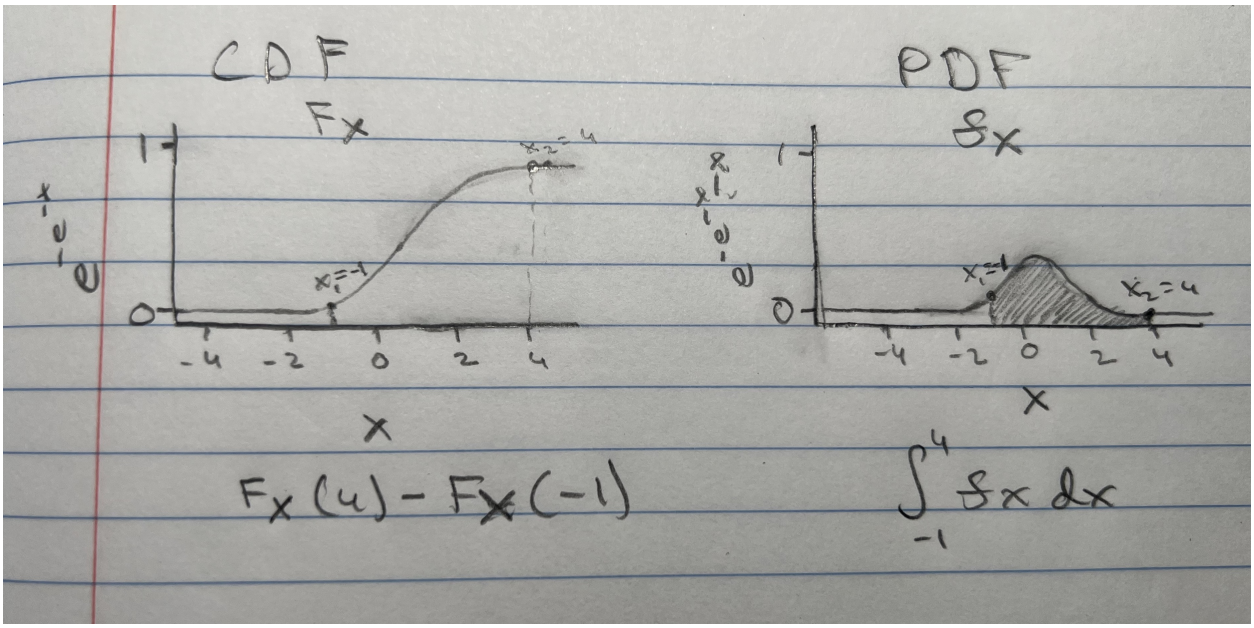


Problem 1 Part e) Sketch the distribution function together with a separate sketch of the density function (side-by-side), indicating $P\{-1 < X \leq 4\}$ on BOTH plots.

Hint: Remember that one will be a difference of y-value (can be indicated as a vertical line segment between the heights at the two x-values) and the other will be an area (can be indicated as shading the positive area under the curve between the two x-values). It's up to you to determine which is which.

Hint: Want a challenge? Try producing these figures in R. (See the Helpful Handout regarding Shading Areas in R.)

NOTE: Upload an image of your sketches.



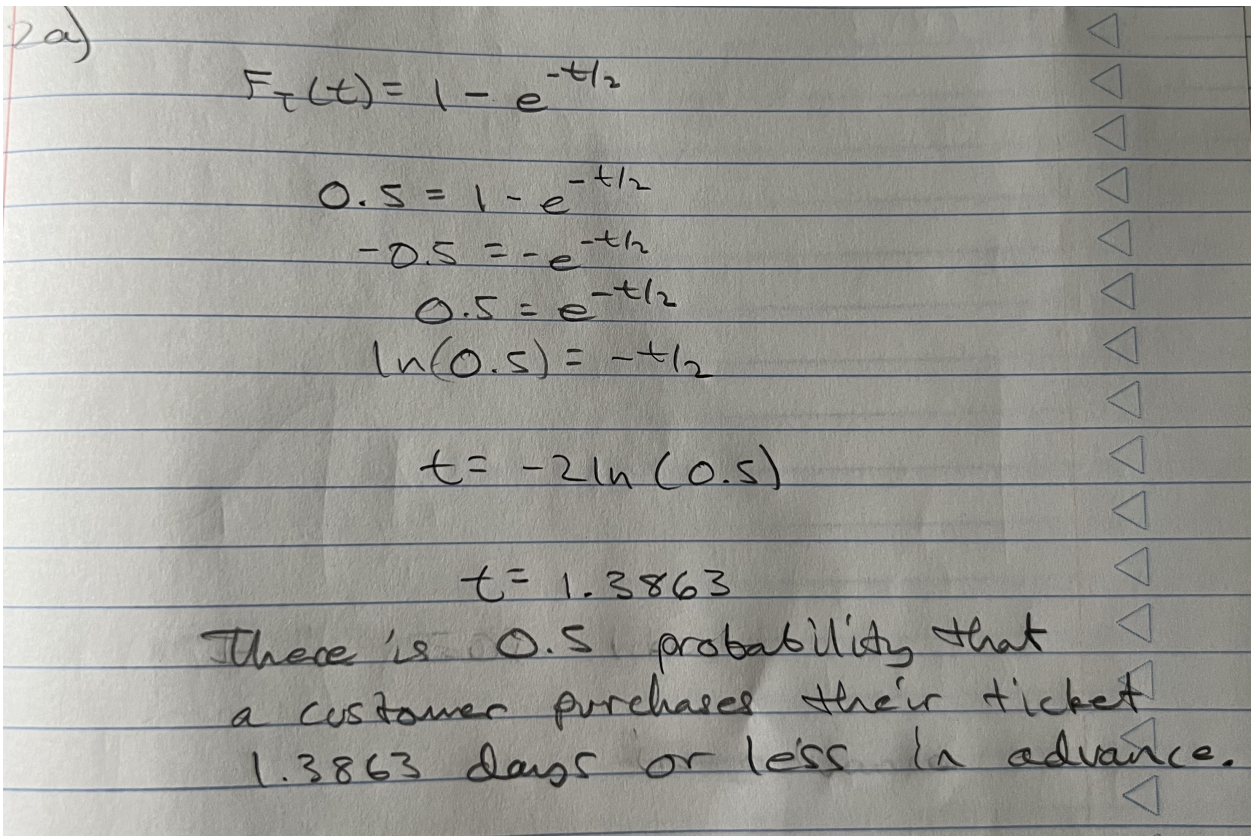
Problem 2. The time spent waiting between events is often modeled using the exponential distribution. For example, suppose that for a given movie, the number of days in advance that customers can purchase their tickets can be modeled by an exponential distribution with an average advance purchase time of 2 days. Let T be the amount of advance purchase time for customers. Then T is a continuous random variable with the cumulative distribution function $F_T(t) = 1 - \exp(-t/2)$.

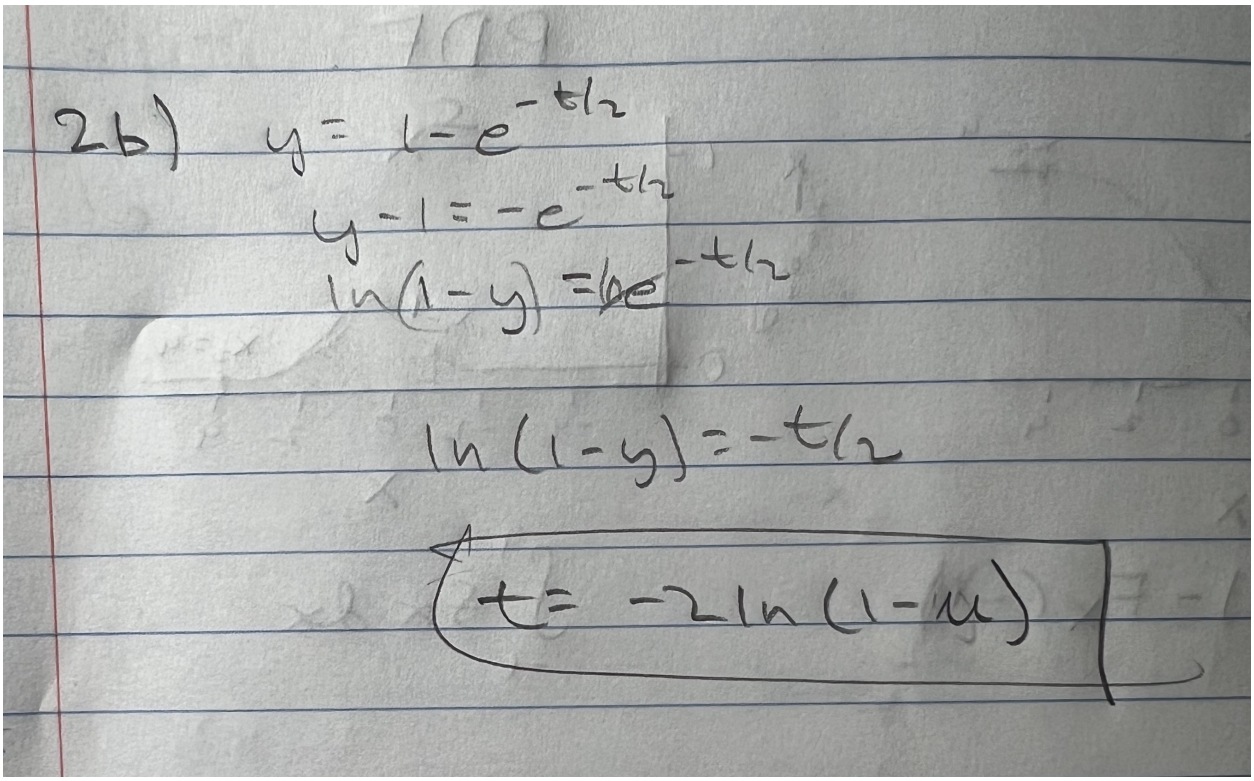
Problem 2 Part a) Determine the median of this distribution, and interpret the value in terms of the context of the scenario provided in the prompt.

NOTE: Upload an image of your work for parts (a) and (b) below part (b).

Problem 2 Part b) Determine the probability transform for this distribution, writing t as a function of u.

Hint: Recall that we can simulate a random variable with distribution $F_T(t)$ through the quantile function $T = F_T^{-1}(U)$.





Problem 2 Part c) Use R and the probability transform you found in (b) to create 80 samples with this distribution. Then do it again, for a total of two simulations of 80 samples with this distribution.

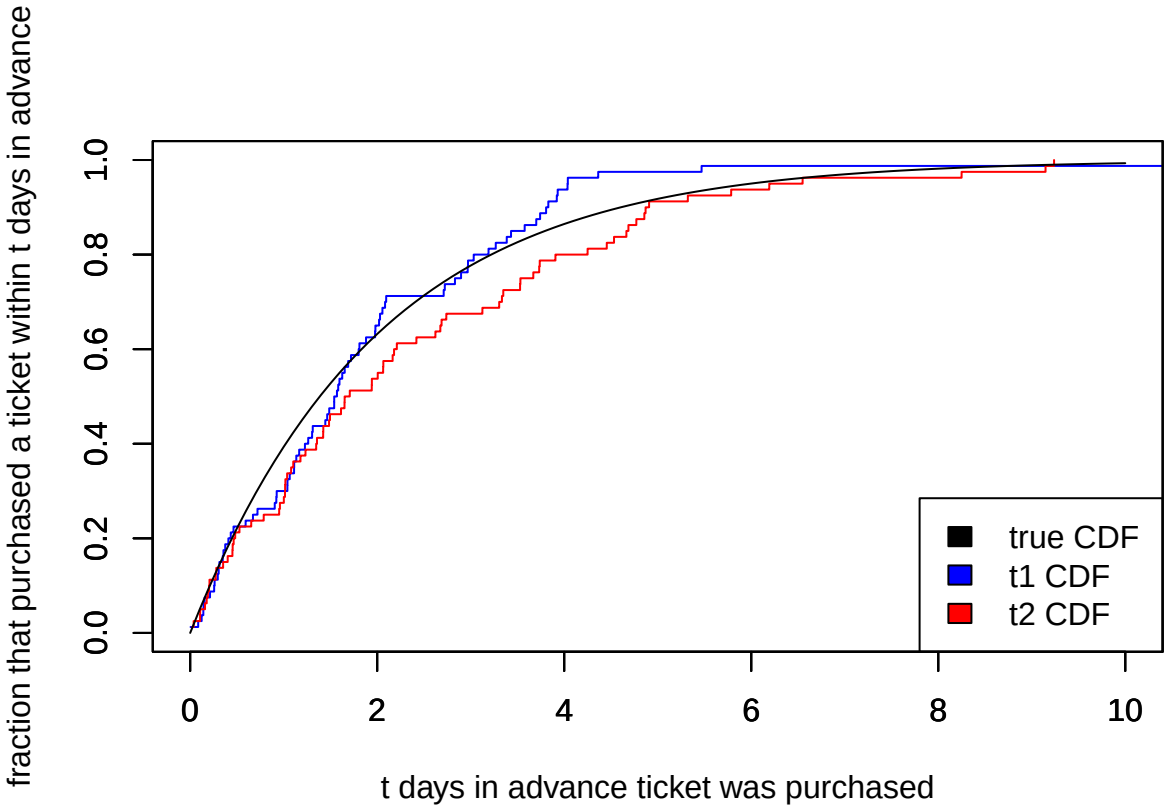
```r
set.seed(2022)  # for repeatability

# NOTE: The line below simulates 80 uniform random variables on [0, 1]
u1 = runif(80, min = 0, max = 1)

# NOTE: The line below finds the quantile function via the probability
# transform NOTE: apply your answer to (b) to u1
t1 = -2 * log(1 - u1)

# NOTE: Now do it again!
u2 = runif(80, min = 0, max = 1)
t2 = -2 * log(1 - u2)
```

Problem 2 Part d) Using R, plot the graph that shows the true cumulative distribution function (in black) along with the empirical CDF of the data (in two clearly visible colors) you simulated in (c). Label the graph in terms of the context of the scenario provided in the prompt and include a legend.

Hint: Recall that to plot the empirical CDF in R, we need to calculate and plot the cumulative probabilities against the sorted data, from smallest to largest. This is similar to the steps taken in Worksheet #02.

\textit{NOTE: To do this task, enter your code in the one of the two R chunks below (and delete the unused chunk).

```r
# NOTE: Do either this chunk or the next one.  You're probably more familiar
# with this style and the other chunk is for the adventurous.

# NOTE: The line below creates 80 equispaced values in (0,1]
y = seq(0.0125, 1, by = 0.0125)


# NOTE: Let's plot the first simulation.
plot(sort(t1), y, type = "s", xlim = c(0, 10), ylim = c(0, 1), xlab = "t days in advance ticket was pur
    ylab = "fraction that purchased a ticket within t days in advance", col = "blue")

# NOTE: Necessary command.
par(new = TRUE)

# NOTE: Now you do the plot of the second simulation.  NOTE: Be extra careful
# that the parameters are set to produce a nice image
plot(sort(t2), y, type = "s", xlim = c(0, 10), ylim = c(0, 1), xlab = "", ylab = "",
    col = "red")

# NOTE: Now we will plot the curve of the true distribution function

# NOTE: Necessary command.
par(new = TRUE)

# NOTE: The command curve() requires its function expression to be in terms of
# x (not t)
curve(1 - exp(-x/2), from = 0, to = 10, ylim = c(0, 1), xlab = "", ylab = "")

# NOTE: Finally, we add a legend.
legend("bottomright", legend = c(" true CDF ", " t1 CDF ", " t2 CDF "), fill = c("black",
    "blue", "red"))  #sets the colors corresponding to the curve names
```



Problem 2 Part e) Describe how well the empirical distributions from your simulated data match the theoretical/true cumulative distribution function, $F_T(t)$. From your graph estimate how many days away from the true distribution median are the medians predicted by the two simulations of 80 data points.

NOTE: Enter your response as text below.

both t1 and t2 empirical distributions match the real CDF fairly well in that they are both close to and resemble the real CDF curve. based on just the graph, I would say the t1 median is about 0.15 days away from the true median, and the t2 median is about 0.3 days away from the true median