DUE DATE: FRIDAY, OCTOBER 22, 2021 by 11:59 PM on Gradescope
LATE DUE DATE: WEDNESDAY, OCTOBER 27, 2021 by 11:59 PM on Gradescope with penalty of 5 percentage points per day late

R: Graphing a Normal Distribution

Directions:

- Submit your project on Gradescope under "R Project 4" under the correct version number. You only need to submit to 1 version (your assigned one). Points will be deducted if the project is not submitted to the correct version.
- Project must be typed for full credit. Write your answers in the R Script (top-left hand corner) area in R Studio. Save your R Script as a .R file. Save any graphs you generate. Submit your .R file and graphs to Gradescope. (More guidance on Pages 2-3)
- Add appropriate amounts of white space to make your response easy to read.
- Report your code for each part, and provide your graphs. Then answer the associated questions. If you type the answers to the questions in your code, clearly indicate which question you are answering.
- Answers without code will not receive full credit. Code without answers will not receive full credit.
- Provide appropriate commentary where needed.
- You may discuss projects with your fellow students (and you may use Piazza), but you must write your answers up independently, and in your own words. Asking for assistance, or viewing solutions, through third-party websites (like Chegg) is not permitted.

Grading Rubric (out of 62 points):

Submitted an R Script File: 1 point

Part 1 (16 points):

- 15 points code
- 1 point graph

Part 2 (33 points):

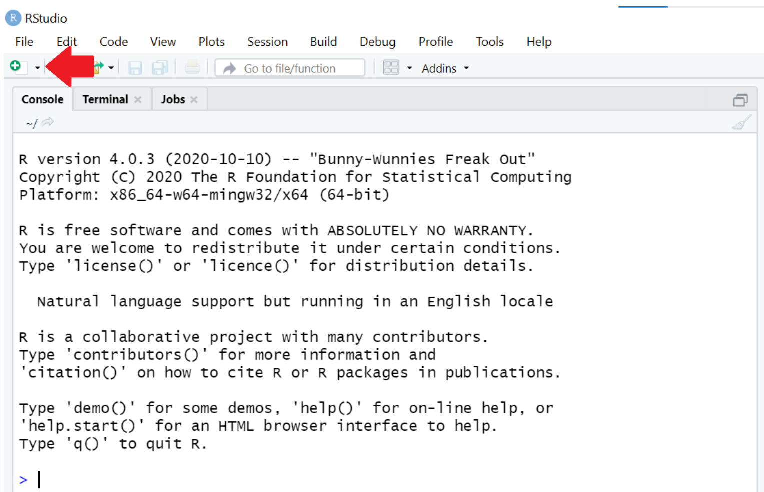
- 32 points code
- 1 point graph

Questions (12 points)

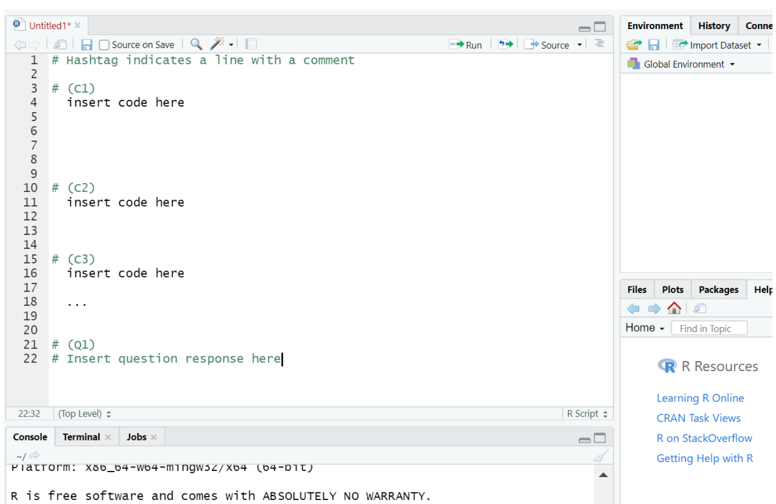
1. 1 point
2. 2 points
3. 3 points
4. 1 point
5. 2 points
6. 2 points
7. 1 point

How to Save your Code and Question Answers:

1. To open a new R Script to type your code, click on the arrow in the figure below. Select "R Script".



2. Once your R Script is open, you can potentially format your file like below. It contains both the necessary code as well as the answers to the questions



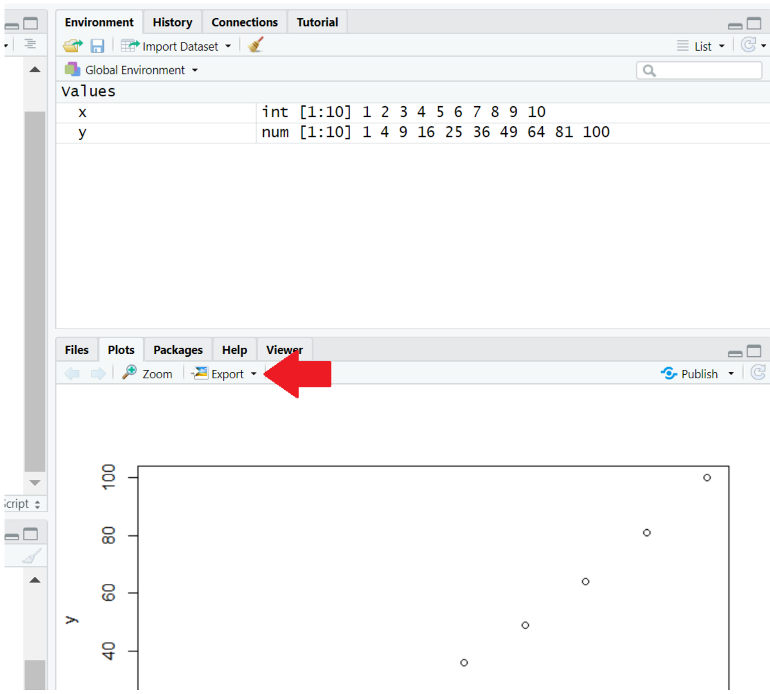
3. When you want to save your file, go to File / Save As... Find where you want to save your file, and save it. You'll have to make sure you put the file extension as .R.

   For Example: "R_Project4.R"

   Save all of your code and answers in 1 file.

How to Save your Graph(s):

1. To save any graphs or plots you may have generated, go to the "Plots" portion of the lower-right-hand corner window. Click on the arrow next to the "Export" button.



2. There are several options to choose from. You may save your graph / plot as an image, or as a PDF. Either is acceptable.

Some R Functions You Will Potentially Need:

1. Basic R Syntax and Saving Numbers and Vectors: See Swirl Module 1

2. Sequences of Numbers: See Swirl Module 3

3. Normal Distribution: Define `mean` $= \mu$ and `sd` $= \sigma$. The default in the code below is when $\mu = 0$ and $\sigma = 1$, or when we have a *standard normal distribution*, $\mathsf{N}(0, 1)$.

    - `dnorm(x, mean, sd)` Gives the density, $f(x)$, for a Normally Distributed Random Variable. `mean` and `sd` are as defined above.

    - `pnorm(q, mean, sd, lower.tail = TRUE)`: Fins $\mathsf{P}(X \leq q)$ for a Normally Distributed Random Variable. `lower.tail=TRUE` represents the probability $\mathsf{P}(X \leq q)$. If you wanted to find $\mathsf{P}(X > q)$, then you would state `lower.tail=FALSE` instead.

    - `qnorm(p, mean, sd, lower.tail = TRUE)`: Takes a probability, p, and outputs the x value associated with it. For example, suppose $X \sim \mathsf{N}(75, \sigma^2 = 25)$. You want to find k such that $\mathsf{P}(X \leq k) = 0.95$. The associated code would be

      ```
      qnorm(0.95, mean = 75, sd = 5, lower.tail = TRUE)
      ```

      The result would be $k = 83.22427$. This is similar to the $\mathsf{invnorm}()$ function on our calculator when we do not convert to a z-score first. In the calculator, we would have $\mathsf{invnorm}(0.95, 75, 5) = 83.22426813$.

      To check our answer:

        - First, find the z-score associated with an area of 0.95 to the left of the value of interest. $\mathsf{invnorm}(0.95) = 1.645$.
        - Next, we need to find the associated value.

          $$z = \frac{k - \mu}{\sigma} \Rightarrow 1.645 = \frac{k - 75}{5}$$
          $$\Rightarrow 1.645(5) + 75 = k \approx 83.225$$

        - Any differences between our answer and the answer from `R` are because we rounded the z-score above, whereas `R` will not round the intermediate calculations.

4. In the above code, you can either replace the variables `x, p, lambda, mean, sd` etc. with the specific value, or type something like `p = 0.95, mean = 75, sd = 5`. This allows you to put the variables in any spot within the function and still obtain the same result. It also helps you remember what number goes with each portion of the function.

5. Plots and Diagrams

- `plot(x, y, type, main, xlab, ylab, pch, col)`:
  - `x`: the x-coordinates of the points in the plot.
  - `y`: the y-coordinates of the points in the plot.
  - `type`: what type of plot should be drawn. Possible types are:
    - p for points
    - l for lines (this is a lower case L)
    - b for both points and lines
    - h for histogram-like vertical lines
    - s for stair steps
  - `main`: an overall title for the plot
  - `xlab`: allows us to change the description of the values on the x-axis. For example, `xlab = "Number of socks"`.
  - `ylab`: allows us to change the description of the values on the y-axis. For example, `ylab = "Frequency of Blood Pressure"`
  - `pch`: allows us to change the type of points that are drawn. Possible options are below.

  - `lty`: allows us to change the type of line that is drawn. Possible options are below.
    - 1 = solid (default)
    - 2 = dashed
    - 3 = dotted
    - 4 = dotdash
    - 5 = longdash
    - 6 = twodash
  - `col`: allows us to change the color of the points or lines of the plot. For example, `col = "red"` gives red; `col = "blue"` gives blue.
  - Example: Make a scatterplot of square points that are red, adding some labels.

```
plot(x, y, type = "p",
     main = "Title",
     xlab = "Socks",
     ylab = "Frequency",
     pch = 0, col = "red")
```

- `abline(a, b, h, v, lty, col)`: This function adds one or more straight lines through the current plot.
  - `a`: the y-intercept of the line you want to draw
  - `b`: the slope of the line you want to draw
  - `h`: if you want to draw a horizontal line, can specify the y-value for the line
  - `v`: if you want to draw a vertical line, can specify the x-value for the line
  - Examples:
    (a) You want to draw the line $y = -6x+2$ on your graph. `abline(a = 2, b = -6)`

    (b) You want to draw the line $y = 8$ on your graph (a horizontal line at $y = 8$). `abline(h = 8)`

    (c) You want to draw the line $x = -3$ on your graph (a vertical line at $x = -3$). `abline(v = -3)`
  - `lty`: allows us to change the line type
    * 0 = blank
    * 1 = solid (default)
    * 2 = dashed
    * 3 = dotted
    * 4 = dotdash
    * 5 = longdash
    * 6 = twodash
  - `col`: allows us to change the color of the points or lines of the plot. For example, `col = "red"` gives red; `col = "blue"` gives blue.
  - Example: Draw a red, vertical, dashed line at $x = -2$.
    `abline(v = -2, col = "red", lty = 2)`

R Project Version 3
Graphing a Normal Distribution

Part 1. Generate a plot of the standard normal density function (PDF).

- **(C 1)** We want the x-values to be a sequence from -3.35 to 3.35 by 0.01. Store these as `xvals1`.

- **(C 2)** To generate the associated y-values from the standard normal distribution, we input the `xvals1` from **(C 1)** into the probability density function from a standard normal distribution. Store these as `yvals1`.

- **(C 3)** Create a plot of the above x-values and y-values. While creating your plot,
  - Your graph should be a continuous line.
  - The line should be the color "khaki4". You MUST use the specific color name that we have provided.
  - Set the title of the plot to be "Standard Normal Density Function".
  - Set the x-axis label to be "Standard Normal Variable".
  - Set the y-axis label to be "Density".

- **(Graph 1)** After you finish **(C 3)**, save your plot. You will submit it when you submit your code.

Part 2. Generate a plot of the Normal Cumulative Distribution Function (CDF), where $X \sim \mathsf{N}(\mu = 450,\, \sigma^2 = 289)$

- **(C 4)** We want the x-values to be a sequence from 390 to 510 by 4. Store these as `xvals2`.

- **(C 5)** To generate the associated y-values from the cumulative standard normal distribution ($\mathsf{P}(X \leq x)$), we input the `xvals2` from **(C 4)** into the (cumulative) probability function from a standard normal distribution. Store these as `yvals2`.

- **(C 6)** Create a plot of the above x-values and y-values. While creating your plot,
  - Your graph should be a continuous line.
  - The line should be the color "darkslategrey". You MUST use the specific color name that we have provided.
  - Set the title of of the plot to be "Normal CDF Function".
  - Set the x-axis label to be "Normal Variable".
  - Set the y-axis label to be "Cumulative Probability".

- **(C 7)** Save the values $0.025, 0.25, 0.50, 0.75, 0.975$ as a single vector named `cumul_pbty`.

- **(C 8)** Overlay *dashed* horizontal lines at the probabilities indicated in **(C 7)**. Use the color "tan1". You MUST use the specific color name that we have provided.

(C 9) We want to find the x-values associated with the cumulative probabilities given in **(C 7)**, meaning, we want to solve $\mathsf{P}(X \leq k)$ for k, where the values of the probability $\mathsf{P}(X \leq k)$ are the values given in **(C 7)**. This is called finding quantiles / percentiles. Save your vector of quantiles as `quantile_values`.

To do this, you'll use the appropriate function from the normal distribution, and for your first input, use the vector you created in **(C 7)**.

(C 10) Overlay *solid* vertical lines at `x = quantile_values`. Use the color "springgreen2". You MUST use the specific color name that we have provided.

(Graph 2) After you finish **(C 10)**, save your plot. You will submit it when you submit your code.

Questions:

(Q 1) What is the (approximate) largest y-value on your Standard Normal Density Function?

(Q 2) Describe the Standard Normal Density Function. Is it symmetric or skewed? If it is symmetric, is it bell-shaped or uniform? If it is skewed, is it left- or right-skewed?

(Q 3) Do the x-values of the tails of the Standard Normal Density Function really stop at ± 3.35? Why or why not? If not, what values do they go to?

(Q 4) Based on your graph in Part 1, what do the x-values represent?

(Q 5) In Part 2, what do the y-values approach on the CDF curve as $x \to \pm\infty$?

(Q 6) In Part 2, what do the overlaid vertical and horizontal lines represent where they intersect the CDF curve?

(Q 7) Using the graph of the Normal CDF, identify the value of the CDF where the normal variable is equal to 416.6806. Hint: One of the horizontal lines also goes to this point.