STAT 471: Homework 1

Due: September 13, 2025 at 11:59pm

1 Instructions

Please make sure to submit your solutions to the following questions in an .rmd file and a knitted .html file format (or you can use an .nb.html if you'd like). For question 2, please show all work.

2 Question 1 (25 points)

In class, we saw how to sort a vector without using any built-in functions using the Bubble Sort algorithm. Bubble Sort is iterative so it can be slow with large vectors, you will write an R function to perform a faster recursive method called Quicksort.

Step 1. Create the function. For the "if condition", if the size of the vector is less than or equal to 1, return the original vector.
Step 2. Now for the "else condition", create a new object and assign the first element of the vector (call it pivot). Create an object called left that compares the elements of the vector to the left of the pivot. Do the same for the middle (equal to pivot) and right objects (to the right of the pivot). These will form partitions of the vector.
Step 3. Return a vector that consists of the quicksort function being called with the left as input, middle, and quicksort function being called with the right as input. This vector should be of length 3.
Step 4. Test your quicksort implementation with a randomized vector. You can use the code provided below after the implementation:

```
set.seed(42)
vec = sample(1:50, 10)
vec

sorted_vec = quicksort(vec)
sorted_vec
```

3 Question 2 (25 points)

Suppose you have the function $f(x) = x^2 - 6$. Perform the first two iterations of Newton's method to find x_1 and x_2 using the initial guess $x_0 = 2.5$. Compare your solution with the actual answer $\sqrt{6}$.

4 Question 3 (25 points)

In machine learning, we traditionally split the data we are working with into two sets; where 80% of the data is allocated into the training set and the remaining 20% is allocated to the testing set. We will be splitting the built-in dataset "ToothGrowth" into training and testing sets using the Tidyverse package, rsample.

(a) Load the Tidyverse and rsample libraries. In your .rmd file, also load the ToothGrowth dataset and assign the dataframe to a new object, call it tooth_data.
(b) Set a random seed (this ensures reproducibility so you get the same results every time you run

your code). Use the initial_split function to split the Toothgrowth data into 80% training and 20% testing sets.
(c) Check the row counts for each of: the original data, the training set, and the testing set. Display the head() of both the training and testing sets.

5 Question 4 (25 points)

Load the Iris dataset in R. Do the following:

(a) Display the head of the dataset using the head() built-in function.
(b) Check the dimensions of the dataset using dim(). Use colnames() to display the column names of the dataset.
(c) Extract the 15th observation from the Sepal.Width column. What is the value of this 15th observation?
(d) Suppose we want to change the 5th observation of the Sepal.Length column to a new value, say 10.6. Make this change to the dataframe. Also change the 1st and 3rd observations of the Sepal.Width column to be both 25.1. Use head() to double-check your work and display your results.