**BIOST 514/517**

**Discussion Section: Week 2**

**Introduction to R**

**R Exercises**

**Exercises:**

1. Download and read in the salary.txt dataset from the “files” folder within the course canvas page, using a function from the command line; call this data object **salary**. Also using a function from the command line, read in the salary.txt dataset which is available on the web at:

<http://faculty.washington.edu/tathornt/Biost509/DataSets/salary.txt>;

call this data object **othersalary**.

Verify that **salary** and **othersalary** are the same. Hint: First use the help page to learn about the ***identical()*** function (e.g., use the “?identical” command) and then use the function.

**Answer the questions below using the salary dataset.**

1. In class today, we saw that 4 entries had missing salary variables. Using the ***is.na()*** function and other commands, determine which number rows have these missing values. (use the “?is.na” command for more information about this function)
2. Create a dataset with the 4 entries (rows) with missing salary data excluded, and rename the data object salary.

For the remaining questions, use the dataset *without* the 4 entries (rows) containing NAs for the salary variable (19788 rows, 1596 unique ids).

1. What is the mean salary of the entries? What proportion of the entries have a greater “salary” than the mean “salary”?
2. What is the earliest “startyr”? What is the “id” of the individual with the earliest start year?
3. Which “rank” has the largest average “salary” (averaged over all entries)? Which “rank” has the lowest average “salary”? What is the difference between of the ranks with the highest and the lowest average salaries?