Lab 4

Statistical Computing, 36-350

Friday September 25, 2015

Today's agenda: Using regular expressions to extract data from text; text manipulations; getting used to very skewed distributions.

General instructions for labs. Upload an R Markdown file, named .Rmd", to Blackboard. You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. Include the name of your lab partner at the top of the file.

R Markdown setup. Open a new R Markdown file; set the output to HTML mode and click "Knit HTML". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission. Alternatively, you can start from the lab's R Markdown file posted on the course website, as a template.

Part I: Rich folks

The file http://www.stat.cmu.edu/~ryantibs/statcomp/labs/rich.html contains a listing of the 100 richest people in America, according to Forbes magazine. We will use the file to practice extracting information from webpages.

1. Use the `readLines()` function to load the file into a character vector called `richhtml`. How many lines does it contain? What is the total number of characters in the file?

2. Open the file in a text editor (*not* as a webpage). Find the entries for Bill Gates and for Elon Musk. Give the text of the lines from the file which record their net worths.

3. Write a regular expression which should capture a person's net worth. Write code, using the `grep()` function, to check that this has exactly 100 matches in `richhtml`, and that the expression is matching the actual net worths (and not just some bit of text associated with them).

4. Write code, using your regular expression from problem 3 and the functions `regexp()` and `regmatches()`, to extract all the net worths from `richhtml`. Check the following:

    a. There are 100 net worths.
    b. The largest net worth is that of Bill Gates, and there is only one person worth that much.
    c. The Koch brothers have the same net worth.
    d. There are 60 people whose net worth is higher than that of Elon Musk, and 4 people whose net worth is the same.

Part II: Spread of wealth

4. The Forbes website writes net worths in the form "\$7,7 B" to mean 7.7×10^9 dollars. Write code to convert from the Forbes format to floating point numbers, and run it to create a vector of net worths, called `networths`. Check the following:

    a. `networths` is indeed a vector, of length 100 and type `double`.
    b. All of the entries in `networths` are greater than 1 billion, and less than 100 billion.
    c. The largest entry in `networths` matches the net worth of Bill Gates.

d. There are 4 entries in `networths` matching the net worth of Elon Musk.

5. Answer the following using the `networths` vector from problem 4.

   a. What is the median net worth of these 100 people?
   b. What is the mean net worth of these 100 people?
   c. How many of these 100 individuals were worth at least 4 billion dollars? 7 billion? 20 billion?

6. Again, answer using the `networths` vector.

   a. What is the total net worth of the 100 richest people?
   b. What fraction of that total is held by the 5 richest people? 10 richest people? 20 richest people?
   c. What is the smallest number of people who together hold at least 75 percent of that total wealth?
   d. Create an empty vector called `fracheld` of length 100. Populate the first entry by the fraction of total wealth held by the richest person (among the 100 total), populate the second entry with the fraction of wealth held by the richest 2 people, the third entry with the fraction of wealth held by the richest 3 people, and so forth. (Hint: you can do this with a for loop. Alternatively, you can use the `cumsum()` function.) Create a line plot of `fracheld` versus the numbers 1 through 100. Check that this line plot visually matches your answers from a through c.
   e. There are about 118 million households in the US, with a total net worth of about 85 trillion dollars (http://www.federalreserve.gov/releases/z1/current/z1.pdf). What fraction of that total wealth is held by the 100 richest people on the Forbes list? What is the ratio of the mean net worth of the richest 100 to the net worth of the mean household?