Lab 9: Tremors

36-350, Statistical Computing

8 November 2013

Agenda: Extracting numbers from text files; change of representation; working with statistical models.

If you have not read the handout for lecture 21 (`http://www.stat.cmu.edu/~cshalizi/statcomp/13/lectures/21/lecture-21.pdf`), now would be a good time to do so.

The file `http://www.stat.cmu.edu/~cshalizi/statcomp/13/lectures/21/ANSS.csv.html` on the class website is a webpage, containing a catalog of all recorded earthquakes from 1 January 2002 through 31 December 2011 (i.e., stopping just before 1 January 2012) which were at least a 6 on the Richter scale. It begins with a bunch of formatting information (or "meta-data"), and then the data itself, one earthquake per line. In this lab, we will parse this data, and fit it to a Poisson model of earthquake intensities.

1. (5) Load the `ANSS.csv.html` file into R. How many lines does it have in total? How many lines actually contain information about earthquakes?

2. (20) Extract the date, time, and magnitude of each earthquake, and store the result in a data frame called `quakes`. Check that it has the correct number of rows and columns. The column for magnitude should contain numeric values, but the other two should contain strings.

   Hint: Section 2.1.2 in the lecture handout.

3. (10) Add a new column to `quakes`, which converts the dates from strings to objects of type `Date`. Check that the conversion is working the way you want it by spot-checking half a dozen random rows.

   Hint: See recipes 7.9, 7.10, 7.11 and 7.13 in *The R Cookbook*.

4. (15) Write code to count how many earthquakes occurred in May 2005, using your `quakes` data frame, and the new column of `Date` values. What is the number? Cross-check this by manually counting in `ANSS.csv.html`.

   Hint: Recipe 7.13.

5. (10) Write a function which takes in an arbitrary month and year, and counts the number of earthquakes during that month. Cross-check this by seeing that it matches your answer for May 2005. Cross-check it again

by seeing that the answer it gives for December 2010 matches what you count by hand.

6. (15) Using your function, create a vector which stores, for every month between January 2002 and December 2011 inclusive, the number of earthquakes which occurred that month. Check that this vector has the proper length (which is what?), and that all of its entries are non-negative whole numbers. Check that it has the proper values for May 2005, and December 2010. If any entries are zero, check in `ANSS.csv.html` that there were in fact no earthquakes during that month.

   Hint: Recipe 7.14 may be helpful, though there are other ways to do it.

7. (5) What is the mean number of earthquakes per month? The median? The standard deviation? The range? (Report all these with only reasonable precision.)

8. (5) Plot a histogram of the number of earthquakes per month, on a probability density (not frequency count) scale. Describe the shape, in words.

9. (10) Suppose we model the number of earthquakes per month as a Poisson distribution. What is the most likely value of the Poisson rate λ? Explain.

   Hint: Lab and homework 7.

10. (5) Add the probability curve for the Poisson to your histogram. How good is the fit?