StatR 101: Fall 2012
Homework 7
Eli Gurarie
Due Wednesday, November 14

Instructions: Please submit a single document with all the R code, short answers and figures. Upload the completed homework assignment into the course webpage drop-box. Don't hesitate to discuss the problems amongst yourselves on the forum. Refer to `lab7.r` for help in manipulating and illustrating distributions.

1. **Halloween Problem:** Three statisticians live on a city block. Being statisticians, they enjoy making Halloween way more complicated and terrifying than it needs to be by distributing candy probabilistically. Mr. Abel has the trick-or-treater roll a six-sided die and gives out as many candies as the number that is rolled (⚀, ⚁, ⚂, ⚃, ⚄, or ⚅). Mrs. Bernoulli has the trick-or-treater flip a coin 6 times, and gives out the total number of heads that turn up. Dr. Cauchy lets trick-or-treaters draw a single card 4 times from a shuffled deck - replacing the card and reshuffling between every draw - and gives a candy out every time that the card is NOT a heart (♡). (Recall that a standard card deck has four suits: hearts - ♡, diamonds - ♢, clubs - ♣, and spades - ♠).

(a) Let X_a, X_b and X_c be the number of candies received at each statistician's home, respectively. Create three vectors `x.a`, `x.b` and `x.c` containing the possible values for these three random variables, and three vectors `f.a`, `f.b` and `f.c` representing their probability mass function.

(b) Plot each of the three probability distributions identified in part 1. Can you name any of these distributions?

(c) Calculate the number of candies that can be expected to be obtained from each of the statisticians $\mathrm{E}(X_k)$ and variances $\mathrm{Var}(X_k)$. You can calculate these using the vectors above, or by hand. From which statistician can a child expect the most candies on average? Which statistician will provide the most consistent (i.e. smallest variance) number of candies?

(d) Let $Y = X_a + X_b + X_c$ represent the total haul of candies. Simulate this process some large number of times (e.g. 10,000) and illustrate the distribution of Y. Confirm that the mean and variance are close to the ones that you predicted above.

(e) Assuming 100 children visit all three homes, use your simulation results to approximate how many children do you expect to get fewer than 5 candies? More than 12 candies? Is this distribution symmetric?

2. **Global Earthquakes I:** There is excellent access to data on global earthquake activity here: http://earthquake.usgs.gov/earthquakes/map/ If you click on "Download Earthquakes" (along the left side of the page) you can obtain a table (in `.csv` format) of the latest earthquakes over 2.5 in magnitude in the past 7 days. We would like to analyze the rate of earthquake occurrence on a global scale. Note that the key column is the "Date" column, and that reading and using Date objects can be tricky. Follow the example code below, based on data that I downloaded on Monday, November 5, 2012 and uploaded to the course website and my faculty page:

```
# the data can be loaded directly from my website:
  earthquakes <- read.csv("http://faculty.washington.edu/eliezg/data/earthquakes.csv")
# Convert the date column to a "date" object in R
  Date <- as.POSIXlt(earthquakes$Date)
# Convert the dates to minutes from the smallest time
  Minute <- as.numeric(Date - min(Date))/60
```

[Note that while this exercise is an almost exact repetition of the volcano analysis in the Week 7 lab.]

(a) Download the latest earthquake data, load the data, obtain the `Minute` vector (as above), and use it to create a vector W representing waiting times (in minutes) between consecutive earthquakes.

(b) What is the (sample) mean and standard deviation of W, the interval between consecutive earthquakes occurring in the past week?

(c) Plot a histogram of W.

(d) Propose a continuous distribution that models these waiting times. Name the distribution and a guess for the value of the key parameter(s). Illustrate this model over a density histogram of waiting times.

(e) What are the assumptions behind your model? Do you think they are appropriate for these data?

(f) Obtain N_{hour}, the number of earthquakes that have occurred in every hour of the past week.

(g) How do you predict this quantity is distributed? Name the distribution a guess for the value of the key parameter(s).

(h) Illustrate the empirical distribution and theoretical prediction for N_{hour}.

3. **The gamma distribution:** We learned several homeworks ago that the sum of two uniform random variables is a triangle distribution. Consider the sum of two independent exponential random variables: $Y = X_1 + X_2$ where X_1 and $X_2 \sim \text{Exp}(\text{mean} = \lambda)$. The distribution of Y is called the *gamma distribution*[1]. Specifically, $Y \sim \text{Gamma}(k = 2, \lambda)$, where k is the shape parameter, representing how many independent exponential r.v.'s were summed, and λ is the *scale parameter* equal to the mean of the exponential r.v.'s. The probability distribution function (pdf) for the $\text{Gamma}(k = 2, \theta)$ distribution is:

$$f(y|2, \lambda) = \frac{y}{\lambda^2} e^{-\frac{y}{\lambda}}; \text{ for } y \geq 0. \tag{1}$$

(a) Write a function for the $\text{Gamma}(k = 2, \lambda)$ distribution called `Gamma2PDF(x, lambda)` and confirm that it is a valid pdf by integrating it from 0 to ∞.

(b) Illustrate the pdf for $\lambda = 1$ and $\lambda = 2$, calculate the means and variances of these two distributions, and illustrate them with vertical lines at the mean $\pm$ 1 s.d.

(c) Simulate two random vectors X_1 and X_2 from an Exp(2) distribution. Plot the histogram of the paired sum of these vectors, and confirm that the resulting distribution is Gamma(2,4) by drawing a curve over the histogram.

(d) Based on the definition of the Gamma distribution and the results from the problem above, propose a model for the expected waiting time for two earthquake events on the globe.

(e) Obtain from the global earthquake data a vector W_2 representing the time between two consecutive events. Plot a histogram of these results and draw the curve of your Gamma distribution model over it using your function. Confirm that is gives the same curve as `R`'s built-in Gamma distribution function (predictably: `dgamma()`)

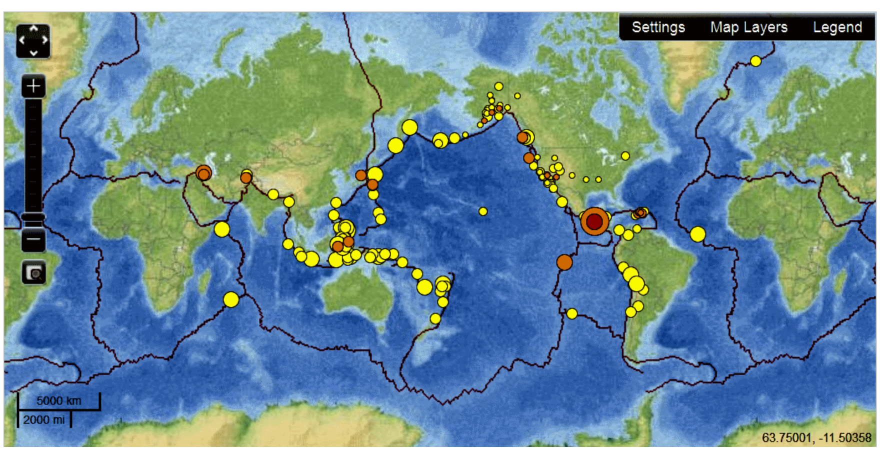

Figure 1: 7 day map of seismic events over 2.5 magnitude, as of noon on Wednesday, November 7, from http://earthquake.usgs.gov/earthquakes/map/.

[1]For more information, see http://en.wikipedia.org/wiki/Gamma_distribution.