Assignment 7

CSCI244-Randomness and Computation
Assigned April 4
Due Friday, April 12

April 4, 2019

One problem about exponential distributions, working with real-life data, another about means, variances, Markov's inequality, and Chebyshev's Inequality.

1 Exponential Distributions in Real Life

'Arrival times', of customers in a queue, of phone calls to a call center, of buses at a bus stop, *etc.,* are frequently modeled by an exponential distribution. (I'm very skeptical about the buses!) In this problem, you will be provided with some real data, posted as a .csv file on the course website. The original dataset, containing the log of 911 calls to emergency services in Montgomery County, Pennsylvania, over a period of several years, was originally posted on Kaggle.com. I have removed some the information that might have served to identify people (!), and limited the calls to those received Monday through Friday from 9 AM to 5 PM, between November 23, 2016, and November 15, 2018, by the emergency medical services. I have also added a few columns to compute the elapsed time, in seconds, between each call and the next. These are the values in the rightmost column of the spreadsheet. There are over 39,000 of these. (The information about the exact time that a call was logged, to the second, was removed, because it was part of a column that contained potentially identifying information; this was used to compute the values in the last column.)

The way I handled 'elapsed time' between the last call logged, say, on November 23 and the first on November 24, was to treat each 9-5 time period as immediately preceding the next: So if the last call on November 23 was at 4:58:20 PM

and the first call on November 24 was at 9:02:53 AM, then the elapsed time is given as 1:40 + 2:53, which is 273 seconds.

There is also code provided for reading the file and extracting the final column, so you can concentrate on the interesting parts.

(a) Write a function that produces a histogram of the data (use the numpy `historgram` function rather than `hist` and use it to produce a stem plot. You should make the size of each bin 60 seconds, so that the histogram will show the number of elapsed times that are less than one minute, between one and two minutes, between two and three minutes, etc. You will probably want to limit the x-coordinates of the display, to get rid of the very long tail where most of the values are 0. The plot should at least *look* like an exponential density.

(b) If the arrival times really are well approximated by an exponential distribution, then the plot should have the form $A \cdot e^{-Bx}$ for some positive constants A and B. How should we determine A and B? (Hint: Look at successive histogram values). Once you've determined these values, superimpose the plot of $A \cdot e^{-Bx}$ on the histogram to see how good a match you get.

(c) If the arrival times really are well approximated by an exponential distribution, then they should exhibit *memorylessness*. This means, for example, that the average waiting time until the next call is the same whether you've just received a call, or whether you've been waiting five minutes for a call. Write a function

```
conditional_mean_wait_time(times,atleast)
```

where `times` is the list of elapsed times, that returns the average elapsed time until the next call, assuming that you have waited at least `atleast` seconds since the preceding call. In other words, you will use only those entries for which the elapsed time `t` is greater than or equal to `atlest`, and compute the average over all these entries of `t-atleast`. What do you find when you set `atleast` to 0 (which gives the mean elapsed time over all entries), 60, 100, 300?

(d) Now write a function that takes the list of elapsed times and a positive integer parameter `k`, and returns a list of times elapsed until the next `k` calls arrive (So, for example, if `k` is 1, you will just get back the original list. If `k` is 2 and the original list of elapsed times is `[14,29,112,78,211]`, you will get back `[43,141,190,289]`.) Plot histograms of this result for `k =2,5,10`. What do you see?

2 Power Law Density

Consider the random variable X whose density function has the form

$$Cx^{-3.1}$$

for $x \geq 1$ and 0 for $x < 1$, where C is some positive constant.

(a) Determine C.

(b) Determine $E(X)$.

(c) Determine var(X).

(d) Chebyshev's inequality gives an upper bound on the probability that the value of a random variable differs by at least two standard deviations from its mean. What is that upper bound? Then verify that X does indeed satisfy this inequality. (Typically, Chebyshev's inequality a very crude upper bound that is far greater than this probability.)

(e) Now consider a random variable Y whose density has the form

$$Cx^{-2.1},$$

for $x \geq 1$, and 0 for $x < 0$, where C is some positive constant. Repeat parts (a,b) above for this density function, and explain why part (c) can *not* be repeated.

(f) Markov's inequality gives an upper bound on the probability that the value of a positive-valued random variable is at least three times its mean. What is that upper bound? Then verify that Y does indeed satisfy this inequality.