Homework 4 (due Friday, November 8, 2024)

1. `(1.5 points)` **Benford's law: Chains of Exponential Distributions**

   As mentioned during the class, many of the chain distributions are Benford. Recall that Expo(λ) is the exponential distribution with parameter λ.

   Check numerically whether Expo(Expo(Expo(Expo(48)))), a chain of 4 exponential distributions indeed satisfies the Benford's law?

2. `(1.5 points)` Based on Theorem 2 of Lectures 11-12, the inter-arrival times are independent and follow the exponential distribution (with parameter λ).

   Let $\{N_t, t \geq 0\}$ be a Poisson point process with rate $\lambda = 2$ on $[0, \infty)$. Write the program to simulate the probabilities of the following problems:

   (a) the probability of two arrivals in $(0, 2]$,

   (b) the probability of two arrivals in $(2, 4]$,

   (c) Compare the answers that you got with the values from the theory.

3. `(1.5 points)` **Poisson Random Variables**

   Let λ_1 and λ_2 be positive real numbers. Suppose that $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$ are independent.

   (a) Prove that $\mathbb{E}(X) = \lambda_1$.
   Hint. Note that

   $$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda_1^k}{k!} \cdot e^{-\lambda_1} = \sum_{k=1}^{\infty} k \cdot \frac{\lambda_1^k}{k!} \cdot e^{-\lambda_1} \\ &= \sum_{k=1}^{\infty} \frac{\lambda_1^k}{(k-1)!} \cdot e^{-\lambda_1} = \sum_{k=0}^{\infty} \frac{\lambda_1^{k+1}}{k!} \cdot e^{-\lambda_1}.\end{aligned}$$

   (b) Prove that $\text{Var}(X) = \lambda_1$. Feel free to use without proof the following formula:

   $$\mathbb{E}(X^2 - X) = \sum_{k=0}^{\infty} (k^2 - k) \cdot \frac{\lambda_1^k}{k!} \cdot e^{-\lambda_1}.$$

   (c) Prove that $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.
   Hint. It suffices to show that for every nonnegative integer a, we have

   $$\mathbb{P}\{X + Y = a\} = \frac{(\lambda_1 + \lambda_2)^a}{a!} \cdot e^{-(\lambda_1 + \lambda_2)}.$$

   Now note that

   $$\mathbb{P}\{X + Y = a\} = \sum_{k=0}^{a} \mathbb{P}\{X = k\} \, \mathbb{P}\{Y = a - k\}.$$

4. `(1 point)` **Exponential Random Variables**

   Let $\lambda > 0$ and $E \sim \text{Expo}(\lambda)$.

   (a) Prove that $\mathbb{E}(E) = \lambda^{-1}$.
   Hint. Recall that

   $$\mathbb{E}(E) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, \mathrm{d}x.$$

   Use integration by parts.

   (b) Prove that $\mathrm{Var}(E) = \lambda^{-2}$.
   Hint. Feel free to use without proof the following formula:

   $$\mathbb{E}(E^2) = \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x}\, \mathrm{d}x.$$

   The integral on the right-hand side can be calculated by integration by parts.

5. `(2 points)` **Approximate Exponential Random Value from Tossing a Physical (Real-life) Coin!**

   In this problem, we would like each student in the class to find a physical coin and individually toss it 74 times. Now before going into the instructions, let us clarify what we mean by *tossing a coin.* First, you can find any coin with two different sides, `HEAD` and `TAIL`. Next, you can design your own way of tossing it. It might take a long time per toss if you spin it and wait until it lands on a face. Instead, we recommend you do something from which you obtain results more quickly. For instance, you can spin it, and once you see it's spinning well, you can hit the coin to make it stop, and read the result.

   The important thing is that you should try to do each toss independently and similarly, so it can be appropriately modeled by *independently and identically distributed* random variables.

   Have a pen and a piece of paper ready, or alternatively, you can use some recording device that you think is appropriate. We also recommend you read through the whole problem first before starting the experiment, so you know what to expect.

   (a) Toss your coin 74 times! Record the results, `H` or `T`, in order. (First row, left to right, then second row, left to right, and so on.) We have provided a table with 74 boxes for you.

Note that it should not take you more than 10 minutes to complete this table. I (Pro) finished recording the 74 results within about 7 minutes.

(b) Note that in the recording table above, we have partitioned the boxes into 37 pairs, so we can easily perform *von Neumann's trick*. Let

$$X_1, X_2, \ldots, X_{37}$$

denote the 37 pairs (in the usual order), so that each X_i is either `HH`, `HT`, `TH`, or `TT`.

Throw away every X_i that is either `HH` or `TT` (let's call these X_i "bad pairs"), and keep every X_i that is either `HT` or `TH` (let's call these X_i "good pairs").

Report how many good pairs there are in your table above.

(c) Suppose that the number of good pairs from part (b) is L. Let

$$X_{i_1}, X_{i_2}, \ldots, X_{i_L}$$

be the good pairs (where $i_1 < i_2 < \cdots < i_L$ are indices where the good pairs happen). Create a binary string

$$b_1 b_2 \cdots b_L$$

of length L in the following way: for each $k \in \{1, 2, \ldots, L\}$, let

$$b_k := \begin{cases} 0 & \text{if } X_{i_k} = \texttt{HT}, \\ 1 & \text{if } X_{i_k} = \texttt{TH}. \end{cases}$$

Report your binary string.

(d) From the binary string

$$b_1 b_2 \cdots b_L,$$

create the number

$$\mathbf{u} := 0.\, b_1 b_2 \cdots b_L\, 1,$$

where we add "0." in front and add "1" to the back. Read this as a binary number so that

$$\mathbf{u} = \frac{b_1}{2^1} + \frac{b_2}{2^2} + \cdots + \frac{b_L}{2^L} + \frac{1}{2^{L+1}}.$$

Compute $\mathbf{u}$. Report your answer either in a fraction of integers, or in decimal representation. Do not round your answer.

(e) Compute $\mathbf{x} = -\log(\mathbf{u})$, where log denotes the natural logarithm. Report your answer. For this part, round your answer to 6 decimal digits after the decimal point.

Here is an example. I (Pro) spun and hit a coin 74 times, and collected these results.

H	T	T	H	H	H	T	H	T	T
T	H	T	H	H	T	H	H	T	T
H	H	H	T	H	T	H	H	H	H
T	T	H	T	T	T	T	T	T	H
T	H	H	T	H	H	T	T	T	T
T	T	T	T	T	T	T	T	H	T
T	T	H	T	H	H	T	T	H	H
H	T	H	H						

I found that there are 15 good pairs, and 22 bad pairs. My binary string is

$$b_1 b_2 \cdots b_{15} = 011110000110000,$$

and thus

$$\mathbf{u} = 0.0111100001100001_2 = \frac{30817}{65536} = 0.4702301025390625.$$

We have

$$\mathbf{x} = -\log\left(\frac{30817}{65536}\right) \approx 0.754533.$$

6. `(1.5 points)` **Spooky Buses for Spooktober.** In this problem, we ask you to simulate the spooky bus paradox which we showed in class. But this time, you can run the code on your own, and get spooked in private.

   The setting is as follows. Suppose that at midnight (0:00) of October 25, which we declare to be time $t = 0$ (the unit is "hour"), the first spooky bus leaves the bus stop. Subsequent buses arrive at the bus stop according to a Poisson point process in $[0, \infty)$ of rate 1 bus per hour. Imagine that on the spooky day, October 31, at 7 pm, which is $t = 163$ if I do the calculation right, you decide to visit this bus stop.

   At 7 pm of that day, you see the bus stop station manager, who is probably a human, and you ask them how long ago the previous bus left, and you add that to your waiting time until the next bus arrives. This sum is the inter-arrival time of the interval containing $t = 163$.

   Let us do the simulation where we have 300 buses. We do 10,000 simulations.

   For each iteration, we record

   (i) avg.inter.time, the average inter-arrival time, which is the average of 299 inter-arrival times, and

   (ii) spooky.time, the inter-arrival time of the interval containing $t = 163$.

   We provide a pseudocode in Algorithm 1. You just have to transform it into a real code and run it.

   Please report the four outputs you obtain.

   Now answer the following (theoretical) questions.

   (a) What are the expectation and the variance of the length of the interval that contains $t = 163$?

   (b) What are the expectation and the variance of the length of the interval between the fifth and the sixth buses?

   (c) What are the expectation and the variance of the length of the interval between the last bus before the end (24:00) of October 29 and the first bus after the beginning (0:00) of October 30?

7. `(extra credit +1 point)` **Poisson and Central Limit Theorem.** Let us fix $\lambda = 4$. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $\mathrm{Pois}(\lambda/n)$ random variables. From Problem 3 above, we learn that the sum

$$X := X_1 + X_2 + \cdots + X_n$$

   is a $\mathrm{Pois}(\lambda)$ random variable, and therefore,

$$\mathbb{E}(X) = \lambda = 4 \text{ and } \mathrm{Var}(X) = \lambda = 4.$$

   Consider the following argument.

Algorithm 1 Spooky Buses

```
N.itr ← 10000                                                ▷ Number of iterations
N.bus ← 300                                  ▷ Number of Buses in each iteration
avg.inter.time ← an empty vector of length N.itr       ▷ average inter-arrival times
spooky.time ← an empty vector of length N.itr            ▷ spooky inter-arrival times
for itr= 1 to N.itr do
    bus.time ← an empty vector of length N.bus                     ▷ bus arrival times
    inter.time ← an empty vector of length N.bus−1                 ▷ bus waiting times
    bus.time[1] ← 0.                                              ▷ First bus at t = 0
    for i = 2 to N.bus do
        u ← runif(1)                                   ▷ Sample u from Unif((0, 1))
        inter.time[i − 1] ← − log(u)                       ▷ Record the waiting time
        bus.time[i] ← bus.time[i − 1] + inter.time[i − 1].
    end for
    avg.inter.time[itr] ← mean(inter.time)                        ▷ Record the mean
    a ← min {i : bus.time[i] > 163}.
    spooky.time[itr] ← bus.time[a] − bus.time[a − 1].  ▷ Record the length of the interval
that contains t = 163
end for
print(mean(avg.inter.time))
print(var(avg.inter.time))
print(mean(spooky.time))
print(var(spooky.time))
```

Note that $X_1, X_2, \ldots, X_n$ are independently and identically distributed. Therefore, the central limit theorem should say that for every pair of real numbers $a \le b$, the probability

$$p_n := \mathbb{P}\left\{a \le \frac{X_1 + X_2 + \cdots + X_n - \lambda}{\sqrt{\lambda}} \le b\right\}$$

converges to

$$\int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t,$$

as $n \to \infty$.

In particular, when $a = 0$ and $b = 1$, this implies

$$\begin{aligned}\lim_{n\to\infty} p_n &= \lim_{n\to\infty} \mathbb{P}\left\{0 \le \frac{X_1 + X_2 + \cdots + X_n - \lambda}{\sqrt{\lambda}} \le 1\right\} \\ &= \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t \approx 0.3413447.\end{aligned}$$

However, recall that $\lambda = 4$ and that $X_1 + X_2 + \cdots + X_n = X \sim \text{Pois}(4)$. Therefore,

$$p_n = \mathbb{P}\left\{0 \le \frac{X - 4}{2} \le 1\right\} = \mathbb{P}\{X \in \{4, 5, 6\}\},$$

and so we can compute

$$p_n = e^{-4} \cdot \frac{4^4}{4!} + e^{-4} \cdot \frac{4^5}{5!} + e^{-4} \cdot \frac{4^6}{6!} \approx 0.4558559,$$

for every n. In particular, p_n does *not* seem to converge to

$$\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d} \approx 0.3413447.$$

What's wrong with this argument?