Chris Piech
CS 109

Problem Set #3
Oct 15, 2018

Problem Set #3 Solutions

With problems by Mehran Sahami and Chris Piech

Warmup

1. Generative simulation:

   a. $X \sim \text{Ber}(p = 0.4)$

```
def simulateBernoulli(p = 0.4):
    if random() < p:
        return 1
    return 0
```

   b. $X \sim \text{Bin}(n = 20, p = 0.4)$

```
def simulateBinomial(n = 20, p = 0.4):
    nSuccesses = 0
    for i in range(n):
        nSuccesses += simulateBernoulli(p)
    return nSuccesses
```

   c. $X \sim \text{Geo}(p = 0.03)$

```
def simulateGeometric():
    nTrials = 0
    while True:
        nTrials += 1
        if random() < 0.03:
            return nTrials
```

d. $X \sim \text{HypGeo}(k = 5, p = 0.03)$

```
def simulateGeometric():
    nSuccesses = 0
    nTrials = 0
    while True:
        nTrials += 1
        if random() < 0.03:
            nSuccesses += 1
        if nSucessess == 5:
            return nTrials
```

e. $X \sim \text{Poi}(\lambda = 3.1)$

```
def simulatePoisson():
    n = 60000
    p = 3.1 / n
    nEvents = 0
    repeat n times:
        if random() < p:
          nEvents += 1
    return nEvents
```

f. $X \sim \text{Exp}(\lambda = 3.1)$

```
def simulateExponential():
    n = 60000
    p = 3.1 / n
    index = 1
    while True:
        if random() < p:
          return index / n
```

2. Lyft line:

a. How much does Lyft expect to make from this trip?

There are three relevant outcomes, there are no requests and the user rides alone (Lyft loses a dollar), there is one request and two users ride in the car (Lyft makes \$5), there are two or more requests and three users ride in the car (lyft makes \$11). Let X be the number of requests that arrive in the next 5 mins. $X \sim \text{Poi}(2)$. Let Y be the number of dollars that lift earns:

$$\begin{aligned}
P(X = 0) &= \frac{2^0}{0!}e^{-2} \\
P(X = 1) &= \frac{2^1}{1!}e^{-2} \\
P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\
&= 1 - e^{-2} - 2e^{-2}
\end{aligned}$$

Now we can calculate the expectation of Y.

$$\begin{aligned}
E[Y] &= (-1)P(Y = 1) + (5)P(Y = 5) + (11)P(Y = 11) \\
&= (-1)P(X = 0) + (5)P(X = 1) + (11)P(X \geq 2) \\
&= -1 \cdot e^{-2} + 5 \cdot \frac{2^1}{1!}e^{-2} + 11[1 - e^{-2} - 2e^{-2}] \\
&= \$7.75
\end{aligned}$$

b. What is the probability that another user will make a request in the next 15 seconds?
The time until the next user can be modelled as an exponential distribution $X \sim \text{Exp}(\lambda = 5)$ however remember that the units of X are “five minutes". In that unit space we want to know $P(X < 0.05)$ (since 15 seconds is 1/20th of five minutes).

$$\begin{aligned}
P(X < 0.05) &= F_X(0.05) \\
&= 1 - e^{-\lambda 0.05} \\
&= 1 - e^{-0.1} \\
&\approx 0.095
\end{aligned}$$

3. Juries:

Let event C = jury is correct. Let event G = defendant is guilty, so G^C = defendant is innocent. Let X = number of jurors that vote guilty, and Y = number of jurors that vote innocent. Note that since the jurors cast their votes independently, the jury's overall decision is determined using a binomial distribution: $X \mid G^C \sim \text{Bin}(12, 0.15), Y \mid G \sim \text{Bin}(12, 0.25)$.

$$\begin{aligned}
P(C) &= P(C \mid G)P(G) + P(C \mid G^C)P(G^C) \\
&= P(Y \leq 3 \mid G)P(G) + P(X \leq 8 \mid G^C)P(G^C) \\
&= \left[\sum_{i=0}^{3}\binom{12}{i}(0.25)^i(0.75)^{12-i}\right](0.7) + \left[\sum_{i=0}^{8}\binom{12}{i}(0.15)^i(0.85)^{12-i}\right](0.3) \approx 0.7541
\end{aligned}$$

Let event F = defendant is convicted (i.e., found guilty by jury). Again the jury's overall decision is determined using a binomial: $Y \mid G \sim \text{Bin}(12, 0.25), Y \mid G^C \sim \text{Bin}(12, 0.85)$.

$$\begin{aligned}
P(F) &= P(F \mid G)P(G) + P(F \mid G^C)P(G^C) \\
&= P(Y \leq 3 \mid G)P(G) + P(Y \leq 3 \mid G^C)P(G^C) \\
&= \left[\sum_{i=0}^{3} \binom{12}{i}(0.25)^i(0.75)^{12-i}\right](0.7) + \left[\sum_{i=0}^{3} \binom{12}{i}(0.85)^i(0.15)^{12-i}\right](0.3) \approx 0.4541
\end{aligned}$$

4. Hash table:

   Let X_i be a binary variable that has the value 1 when there is at least one string hashed to bucket i after the n strings are added to the table (and 0 otherwise). We want to compute:

$$\begin{aligned}
E\left[\sum_{i=1}^{k} X_i\right] &= \sum_{i=1}^{k} E\left[X_i\right] \\
&= \sum_{i=1}^{k} P(X_i = 1) \\
&= \sum_{i=1}^{k} (1 - (1 - p_i)^n) \\
&= k - \sum_{i=1}^{k} (1 - p_i)^n
\end{aligned}$$

5. Let X be a continuous random variable with probability density function:

$$f(x) = \begin{cases} c(2 - 2x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   a. What is the value of c?

      We know $f(x)$ is a PDF, so the area under it must be equal to 1.

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x)dx \\
&= \int_{-1}^{1} \left[c * (2 - 2x^2)\right] dx \\
&= c * \left[2x - \frac{2}{3}x^3\right]_{-1}^{1} \\
&= c * \frac{8}{3} => c = \frac{3}{8}
\end{aligned}$$

b. What is the cumulative distribution function (CDF) of X?

We can use the following relationship to find the CDF of X using $f(x)$:

$$\begin{aligned}
F_X(x) = P(X <= x) &= \int_{-\infty}^{x} f(\widetilde{x})d\widetilde{x} \\
&= \begin{cases} 0 & x < -1 \\ \int_{-1}^{x} \left[\frac{3}{8} * (2 - 2\widetilde{x^2})\right] d\widetilde{x} & -1 < x < 1 \\ 1 & x > 1 \end{cases} \\
&= \begin{cases} 0 & x < -1 \\ \frac{3}{8} \left[*(2\widetilde{x} - \frac{2}{3}\widetilde{x^3})\right]_{-1}^{x} & -1 < x < 1 \\ 1 & x > 1 \end{cases} \\
&= \begin{cases} 0 & x < -1 \\ \frac{1}{2} + \frac{3}{4}x - \frac{1}{4}x^3 & -1 < x < 1 \\ 1 & x > 1 \end{cases}
\end{aligned}$$

c. What is $E[X]$?

A quick way to do it is to notice that $f(x)$ has even symmetry, and thus can be "balanced" around $x = 0$. So $E[X] = 0$.

Another way is to apply the definition of expectation:

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x * f(x)dx \\
&= \int_{-1}^{1} \left[\frac{3}{8} * (2x - 2x^3)\right] dx \\
&= \frac{3}{8} * \left[x^2 - \frac{1}{2}x^4\right]_{-1}^{1} \\
&= 0.
\end{aligned}$$

6. SAT.

a. What fraction of students receive a score within one standard deviation of the mean? Let X be the score of the student. $X \sim N(\mu = 500, \sigma^2 = 100^2)$

$$\begin{aligned}
P(400 < X < 600) &= F_X(600) - F_X(400) \\
&= \Phi(\frac{600 - \mu}{\sigma}) - \Phi(\frac{400 - \mu}{\sigma}) \\
&= \Phi(\frac{600 - 500}{100}) - \Phi(\frac{400 - 500}{100}) \\
&= 2 \cdot \Phi(1) \approx 0.67
\end{aligned}$$

b. Irina scores 750. What percent of students scored lower than 750? (Irina's percentile)

$$\begin{aligned}\text{Percent lower} &= 100 \cdot P(X < 750)\\ &= 100 \cdot \Phi(\frac{750-500}{100})\\ &= 100 \cdot \Phi(2.5) \approx 99.4\%\end{aligned}$$

7. Let X be a Normal random variable with $\mu = 6$. If $P(X > 9) = 0.3$, what is the approximate value of Var(X)?

$$\begin{aligned}0.3 &= P(X > 9)\\ &= 1.0 - F_X(9)\\ &= 1.0 - \Phi\Big(\frac{9-\mu}{\sigma}\Big)\\ 0.7 &= \Phi\Big(\frac{3}{\sigma}\Big)\\ \Phi^{-1}(0.7) &= \frac{3}{\sigma}\\ 0.52 &\approx \frac{3}{\sigma} && \text{In the Phi table, find the value that produces 0.7}\\ \sigma &\approx 5.77\\ Var(X) &= \sigma^2 \approx 33.3\end{aligned}$$

8. The Huffmeister floodplane in Houston has historically been estimated to flood at an average rate of 1 flood every 500 years. A flood plane with that rate of flooding is called a "500 year" floodplane.

a. What is the probability of observing at least 3 floods in 500 years?
Poisson RV with $\lambda = 1$ (flood per 500-year period)

$$\begin{aligned}P(X \geq 3) &= 1 - \sum_{i=0}^{2} P(X = i)\\ &= 1 - \sum_{i=0}^{2} \frac{\lambda^i}{i!} e^{-\lambda}\\ &= 1 - \frac{5}{2e}\end{aligned}$$

b. What is the probability that a flood will occur within the next 100 years?

$$Y \sim \text{Exp}\left(\frac{1}{500}\right) \qquad F_Y(y) = 1 - e^{-\lambda y}$$

$$P(Y < 100) = F_Y(100) = 1 - e^{-\frac{100}{500}} = 1 - e^{-\frac{1}{5}}$$

c. What is the expected number of years until the next flood?
   Using $\lambda = \frac{1}{500}$ in terms of one year

$$E[Y] = \frac{1}{\lambda} = 500$$

d. 10 independent 500 year floodplanes:
   First find the probability p^* that a single floodplane has at least 3 floods:
   $X \sim \text{Poi}(\frac{1}{500})$ = number of floods in a floodplane in one year

$$p^* = P(X \geq 3) = 1 - \sum_{i=0}^{2} \frac{\lambda^i}{i!} e^{-\lambda}$$

$$p^* = 1 - \sum_{i=0}^{2} \frac{(\frac{1}{500})^i}{i!} e^{-\frac{1}{500}}$$

   Then treat each floodplane as an independent trial of a binomial RV with $n = 10$ and $p = p^*$:

   $Y \sim \text{Bin}(10, p^*)$ = number of floodplanes with $X \geq 3$

$$P(Y > 2) = \sum_{i=3}^{10} \binom{10}{i} (p^*)^i (1 - p^*)^{10-i}$$

9. Hindenbug

a. What is the probability that the bug manifests for a given user?
   Let X be the amount of time, in hours, until the bug occurs. $X \sim \text{Exponential}(\lambda = 1/200)$.
   Let E be the even that a bug manifests for the user.

$$P(E) = P(X < 3) = 1 - e^{-\frac{3}{200}} \approx 0.015$$

   Alternatively, you can model the probability as $P(Y > 0)$ with Y being the number of times the bug occurs in 3 hours.

b. Probability that more than 10000 users experience the bug.
   Let $p = 0.015$ be the solution to part (a).
   Let X be the number of users who experience the bug. $X \sim \text{Bernoulli}(n = 10000, p)$
   Let Y be a Normal approximation of X. $Y \sim N(\mu = 10^6 p, \sigma^2 = 10^6 p(1 - p))$.

$$\begin{aligned} P(X > 10000) &= 1 - P(X \leq 10000) \\ &\approx 1 - P(Y < 10000.5) \\ &\approx 1 - \Phi\Big(\frac{10000.5 - 10^6 p}{\sqrt{10^6 p(1 - p)}}\Big) \end{aligned}$$

Dithering

10. The first thing to check is that both sequences produce roughly 150 heads...which is true. Alas, no easy answer. Okay, well, next we notice that the length of runs is substantially different. Notice that in sequence 2 there is a run of 7 tails and later a run of 7 heads. In the first sequence there are only one or two sequences that are of length 4. Which one is better? If the coin flips are independent, then you would expect the distribution of runs to be Geometric. Let X be the length of a run. $X \sim$ Geo(0.5). If you calculate the number of lengths of runs in each sequence, you can see that the PMF from sequence 2 looks much more similar to the Geometric PMF than the PMF from sequence 1. Another way of answering the question is to look at the probability of getting a tails after a sequence of three heads. In sequence 1 the probability is 1/18 = 0.06, when it should be around 0.5. If we model the number of run extensions after 3 as R ~Bin(n = 18, p = 0.5), the probability of getting 1 (or fewer) extensions is 0.5 18 + 18 Âů 0.5 18 < 0.00007. Very suspicious.

Analysis of Bloom Filters

11. We can think each whether each string gets hashed to the first bucket or not as a series of Bernoulli trials (one per string) or a Binomial distribution (with $n = 1,000 * 3$). Since n is large ($n = 3,000$) and p is small ($p = 1/9,000$), a Poisson approximation is perfectly reasonable to use. For parts (a) and (b) we use a Poisson approximation with $\lambda = 3,000/9,000 = 1/3$.

    a. $X \sim \text{Poi}(\lambda = 1/3)$, we compute: $P(X = 0) = e^{-1/3} \approx 0.717$
    b. $X \sim \text{Poi}(\lambda = 1/3)$, we compute: $P(X < 10) \approx 1.000$
    c. We make two large assumptions here. Since it is truly unlikely that two hash functions produce the same bucket, we assume that does not happen (taking it into account won't change our estimate). Similarly the problem lets us assume that the value of one bucket is independent of another (again taking into account the non-independence wouldn't affect the probability).

       Let M be the event that the string is misclassified as "in" the set. Let E_i be the event that there is a 1 in bucket i. By symmetry $P(E_i) = P(E_0)$ where $P(E_0)$.

       $$\begin{aligned} P(M) &= P(E_i)^3 \\ &= [1 - P(E_i^C)]^3 \\ &\approx [1 - 0.717]^3 \approx 0.023 \end{aligned}$$

    d. If there were only 1 hash function, then the number of hashes into the bloom filter would be fewer (only 1,000 as opposed to 3,000) but when looking up a string not in the set, if there happens to be a 1 in that single "query" bucket, you will misclassify the string.

       We can think each how many of the 1,000 original strings get hashed to the query bucket as a Binomial distribution (with $n = 1,000$, $p = 1/9,000$). Since n is large and p is small, a Poisson approximation is perfectly reasonable to use. We use a Poisson approximation with $\lambda = 31,000/9,000 = 1/9$.

Let X be the poisson random variable that approximates the number of strings hashed to the query bucket Let E be the event of a misclassification: at least one string in the original 1,000 is hashed to the query bucket.

$$P(E) = 1 - P(X = 0) = 1 - e^{-\lambda} = 1 - e^{-1/9} \approx 0.105$$

Using three hash functions is a much better choice!

Climate Change

12. a. Estimate the probability that Climate Sensitivity is greater than 7.5 degrees Celsius.

The probability that S < 7.5 is 0.94 (by adding all the values in the PMF). Thus the probability that S > 7.5 is 0.06

b. Calculate the value of K for both f1 and f2.

For f_1

$$\begin{aligned}
0.06 &= \int_{7.5}^{30} \frac{K}{x} \partial x \\
&= K \int_{7.5}^{30} \frac{1}{x} \partial x \\
&= Klog(x)\Big|_{7.5}^{30} \\
&= K(\log(30) - \log(7.5)) \\
K &\approx 0.1
\end{aligned}$$

For f_2

$$\begin{aligned}
0.06 &= \int_{7.5}^{30} \frac{K}{x^3} \partial x \\
&= K \int_{7.5}^{30} \frac{1}{x^3} \partial x \\
&= -K\frac{1}{2x^2}\Big|_{7.5}^{30} \\
&= -K(\frac{1}{1800} - \frac{1}{112.5}) \\
K &\approx 7.2
\end{aligned}$$

c. Estimate the probability that S is greater than 10 under both the f_1 and f_2 assumptions.

Under assumption 1

$$\begin{aligned} P(X > 10) &= \int_{10}^{30} f_1(x)\partial x \\ &= \int_{10}^{30} \frac{1}{10x}\partial x \\ &= \frac{1}{10}\log(x)\Big|_{10}^{30} \approx 0.05 \end{aligned}$$

Under assumption 2

$$\begin{aligned} P(X > 10) &= \int_{10}^{30} f_1(x)\partial x \\ &= \int_{10}^{30} \frac{7.2}{x^3}\partial x \\ &= -\frac{7.2}{2}\cdot\frac{1}{x^2}\Big|_{10}^{30} \approx 0.03 \end{aligned}$$

d. Calculate the expectation of S under both the f_1 and f_2 assumptions.

The expectation of S, for values under 7.5 is 3.02 degrees
Under assumption 1

$$\begin{aligned} E[X|S > 7.5] &= 3.02 + \int_{7.5}^{30} x\cdot f_1(x)\partial x \\ &= 3.02 + \int_{7.5}^{30} \frac{1}{10}\partial x \\ &= 3.02 + \frac{1}{10}\cdot x\Big|_{7.5}^{30} \approx 5.25 \end{aligned}$$

Under assumption 2

$$\begin{aligned} E[X|S > 7.5] &= 3.02 + \int_{7.5}^{30} x\cdot f_2(x)\partial x \\ &= 3.02 + \int_{7.5}^{30} \frac{7.2}{x^2}\partial x \\ &= 3.02 + -7.2\cdot\frac{1}{x}\Big|_{7.5}^{30} \approx 3.74 \end{aligned}$$

e. Let $R = S^2$ be an approximation of the cost to society that results from S. Calculate $E[R]$.

The expectation of R, for values under 7.5 is 12.06 degrees

Under assumption 1

$$\begin{aligned}
E[R|S > 7.5] &= 12.06 + \int_{7.5}^{30} x^2 \cdot f_1(x) \partial x \\
&= 12.06 + \int_{7.5}^{30} \frac{x}{10} \partial x \\
&= 12.06 + \frac{1}{10} \cdot x^2 \Big|_{7.5}^{30} \approx 54.25
\end{aligned}$$

Under assumption 2

$$\begin{aligned}
E[R|S > 7.5] &= 12.06 + \int_{7.5}^{30} x^2 \cdot f_2(x) \partial x \\
&= 12.06 + \int_{7.5}^{30} \frac{7.2}{x} \partial x \\
&= 12.06 + 7.2 \cdot \log(x) \Big|_{7.5}^{30} \approx 16.39
\end{aligned}$$