Deep Dive 2

April 18, 2022

Problem 1 : Bus Lines

1. Let A be a random variable for the Company A's bus arival time, and B - for company B. Then, let $A \sim Unif(a = 0, b = 10)$ be the arival time for the next Company A bus because we have no information on when Maria arrives at the station and how long it is left for Company A bus to arrive (can be 1 min, can be 10). It means that it can equally likely arrive any time in the next 10 minutes. $B \sim Expo(\lambda = \frac{1}{10})$ will be the arrival time of the next Company B bus. To find $P(B < A)$ - probability that the company B bus arrives faster than company A bus, we can apply the law of total probabilties for continuous random variables by integrating over all cases when B<A and finding the area under the curve, which will give us the corresponding probability. The bounds for the integral are determined by company A bus arrival time : since we want B to be less that A, it should be in the range from 0 to 10 minutes. If it were bigger than 10, company A bus would already have arrived, and A would be bigger than B. 0 is the lower bound because the waiting time can not be negative. Thus, we can apply the formula of the continuous law of total probabiltiies:

$P(B < A) = \int_0^{10} P(B < A|A = a) \cdot f_A(a)da$

The first term in the function states the probability of B<A given that A takes a certain value. This is a definition of a CDF (of exponential distribution in our case). The second term is a PDF of a uniform distribution that models the arrival of the next company A bus.

$P(B < A) = \int_0^{10} P(B < A|A = a) \cdot f_A(a)da = \int_0^{10} P(B < a|A = a) \cdot \frac{1}{b-a}da = \frac{1}{10}\int_0^{10}(1 - e^{\frac{-a}{10}})da = \frac{1}{10}((10 + 10e^{\frac{-10}{10}}) - (0 + 10e^{\frac{0}{10}})) = \frac{1}{e} = 0.368$

2. To find PDF, we will start by finding the CDF first. We assign a random variable T to represent Maria's waiting time for the bus. She will get on any bus that comes first, meaning her waiting time is determined by the time it takes for a quicker bus to arrive. It can be represented as $T = min(A, B)$. By definition, CDF is equal to $P(T \leq t)$. To find this probability, we first can find $P(T > t)$, and then its complement :

$P(T > t) = P(A > t, B > t) = P(A > t)P(B > t)$

We can explain this equation by saying that waiting time T is greater than t only when both buses come later

```
[41]: #Problem 1
import scipy.stats as sts
import numpy as np
import matplotlib.pyplot as plt

trials = 1000000
#generate certain number of random variables from their respective distributions
a = sts.uniform(0, 10).rvs(trials)
b = sts.expon(scale=10).rvs(trials)
#counter and a list to store values
firsts = 0
bus_arrivals = []
#for each trial determine which bus comes first and store the result, count the number of times bus B comes first
for i in range(trials):
    if a[i]>b[i]:
        firsts += 1
        bus_arrivals.append(b[i])
    else:
        bus_arrivals.append(a[i])

print("Experimental probability of Company B's Bus arriving first:", firsts/trials)
print("Theoretical probability of Company B's Bus arriving first:", 0.368)

#plot experimental and theoretical data
mins = np.arange(0.5, 10.5, 1)
probs = [np.exp(-i/10)*(1/5-i/100) for i in x]
plt.hist(t, density=True, label='Experimental PDF', edgecolor='white', alpha = 0.6)
plt.plot(mins, probs, label='Theoretical PDF', color='red')
plt.title("Maria's Waiting Time After Arriving at the Bus Stop")
plt.xlabel('Minutes after Maria arrives at stop')
plt.ylabel('Probability')
plt.legend()
plt.show()
```

```
Experimental probability of Company B's Bus arriving first: 0.368044
Theoretical probability of Company B's Bus arriving first: 0.368
```

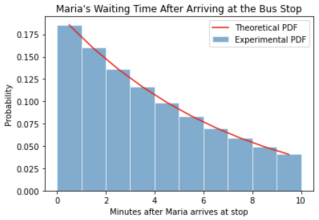

Figure 1: Simulation for Part 1 and 2

than t, so A>t and B>t. We can write it as a multiplication of individual probabilities because A and B are independent random variable : arrival of one bus does not tell us anything or impact the arrival of the other. So, the CDF is :

$P(T \leq t) = 1 - P(A > t)P(B > t)$

Just like we expressed CDF as a compliment, we can substitute complements of CDFs (uniform for A and exponential for B) to express $P(A > t)P(B > t)$:

$P(T \leq t) = 1 - (1 - F_A(t)) \cdot (1 - F_B(t)) = 1 - (1 - \frac{t-a}{b-a})(1 - 1 - e^{\frac{-t}{10}}) = 1 - e^{-\frac{t}{10}}(1 - \frac{t}{10})$

To find the PDF, we simply take derivative of the CDF:

$f_T(t) = \frac{d}{dt}F_T(t) = \frac{d}{dt}1 - e^{-\frac{t}{10}}(1 - \frac{t}{10}) = -\frac{(t-20)e^{-\frac{t}{10}}}{100}$ for $0 < t < 10$, which is the support of the distribution because Marias waiting time can not be bigger than 10 (at least one of the buses would come by then) and less than 0.

3. We cann see from the Figure 1 that the simulated probability that bus B comes first is the sme as we calculated above, and the plot shows that simulated PDF aligns with the trend of the histogram of the calculated PDF.

Problem 2 : Counting Votes

1. We start by defining variables:

X - number of votes for

Y- number of votes against

D=X-Y - difference in votes for and against

N - number of people arrived (and also the total number of votes) = X+Y $\sim Poisson(\lambda)$

Conditional on the total number of votes, N=n, the votes are independent trials with binary outcome and constant probability p. It makes it :

$X|N = n : Binomial(n, p)$

$Y|N = n : Binomial(n, 1 - p)$

We need to condition on the total number of votes because we do not know what the total number is, and we only have their distribution.

Now we need to find the probability that X=x and Y=y. To do so, we start by applying the law of total probabilities for discrete variables and sum them for all possible values of N:

$P(X = x, Y = y) = \sum_{n=0}^{inf} P(X = x, Y = y|N = n) \cdot P(N = n)$

Since we defined N as X+Y above, we can substitute n as x+y. Implementing this condition allows us to remove the sum because we eliminate all 0 probabilities (for example, when n=11 and x=y=5). Also, since X and Y have the exact same distributions, we can eliminate one of them to avoid redundancy:

$P(X = x, Y = y) = P(X = x|N = x + y) \cdot P(N = x + y)$

The last step is to substitute the PMFs of distributions (Binomial for the first term and Poisson for the second as defined above):

$P(X = x, Y = y) = \binom{x+y}{x} p^x (1 - p)^y \cdot e^{-\lambda} \frac{\lambda^{x+y}}{(x+y)!} = e^{-\lambda p} \frac{(\lambda p)^x}{x!} \cdot e^{-\lambda(1-p)} \frac{((\lambda(1-p))^y}{y!}$

Now we can see that the equation simplifies into the product of 2 Poisson distributions. This shows us that the variables X and Y are independent and we can define the distributions of X and Y with appropriate parameters :

$P(X = x, Y = y) = P(X = x)P(Y = y)$

$X \sim Poisson(\lambda p)$

$Y \sim Poisson(\lambda(1 - p))$

Finally, we can find the expected value of D using hte rule of linearity of expectations. We can look up that the expected values of Poisson variables are simply their rate parameters:

$E(D) = E(X - y) = E(X) - E(Y) = \lambda p - \lambda(1 - p) = \lambda(2p - 1)$

For varience, we use the known formula:

$Var(D) = Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$

$Corr(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$ - because the variables are independent, there is no correlation between them

$Var(D) = Var(X - Y) = Var(X) + Var(Y) = \lambda p + \lambda(1 - p) = \lambda$ (we also look up the varience for Poisson variables, and they are also equal to the rate parameters)

Figure 2: Simulation for Problem 2

2. From the Figure 2 we see that I varrief both p and lambda and showed with 2 graphs that the simulated mean is the same as the theoretical. The same is true for the varience in the third graph.

Problem 3 : Hereditary Heights

1. We can consider each child independently first. If we take 1 child and 2 parents, the probability that this child being the tallest is simply $\frac{1}{3}$. Now, since the heights are independent and they all follow the same distribution, each child has the same probability of being taller than both parents and we can multiply the porbability we found by the total number of children to get the expected value (average number of children taller than both parents): $4 \cdot \frac{1}{3} = \frac{4}{3}$.

2. Given that mother's and children's heights are defined by normal distribution, their difference is also defined by normal distribution. This differences will be denoted as random variable D:

$D = Y_j - X_1$ - diffrence between a child's and mother's heights.

Now we need to find the mean and varience of the distribution of D. To find the mean, we simplu subtract means of Y and X distributions :

$u_{Y_j - X_1} = u_{Y_j} - u_{X_1} = 0$ - because the marginal distributions are the same for both variables.

Using the formula for standrad deviation of the distribution of difference between 2 normal random variables (incorporates correlation coefficient):

$$\sigma_{Y_j - X_1} = \sqrt{\sigma^2_{Y_j} + \sigma^2_{X_1} - 2\rho\sigma_{Y_j}\sigma_{X_1}}$$

Again, since marginal distributions are the same, the variences are the same, which lets us simplify it further to :

$$\sigma_{Y_j - X_1} = \sqrt{\sigma^2_{Y_j} + \sigma^2_{X_1} - 2\rho\sigma_{Y_j}\sigma_{X_1}} = \sqrt{2\sigma^2 - 2\rho\sigma^2} = \sigma\sqrt{2(1-p)}$$

Thus, we get the following distribution of the difference in heights:

$Y_j - X_1 = D \sim Normal(0, (\sigma\sqrt{2(1-p)})^2)$

To find the probability that a child is at least 1 cm taller than mother, we can start by standardizing the normal distribution (transforming to standard normal distribution CDF) of height differences using the appropriate general transformation formula:

$Z = \frac{X-\mu}{\sigma}$

In temrs of probabilities, we use the compliment of the probability that the child is less than 1 cm taller than mother, which is the definition of the CDF that we can express using the formula above:

$P(D \geq 1) = 1 - P(D \leq 1) = 1 - \Phi(\frac{x-\mu_D}{\sigma_D}) = 1 - \Phi(\frac{1}{\sigma\sqrt{2(1-p)}})$

To help us find how many children are taller than the mother, on average, we can define an indicator variable I_j - jth child is at least 1 cm taller than mother. We also know that $E(I_j) = P(D \geq 1)$, or $E(I_j) = 1 - \Phi(\frac{1}{\sigma\sqrt{2(1-p)}})$.

Finnaly, we can use the linearity of expectation property and multiply this expected value by 4 (for 4 children). We can apply it because marginal distributions and corelations with mother for all 4 children are identical.

Problem 4 : Radioactive Decay

1. We start by following the frequentust aproach and finding the Maximum Likelihood Estimate, which would be our 'best' estimate. We can simply look up the MLE for exponential distribution. It happens to be $\lambda_{MLE} = \frac{n}{\sum_i x_i} = \frac{6}{1.5+2+3.1+4.2+5.1+11.9} = 0.216$.

2. To find the interval, we need to find the lower and upper bounds. We are going to use the formulas for the bounds for exponential distribution confidence interval :

$\lambda_{lower} = \lambda_{MLE}(1 - \frac{1.96}{\sqrt{n}}) = 0.216(1 - \frac{1.96}{\sqrt{6}}) = 0.0431$

$\lambda_{upper} = \lambda_{MLE}(1 + \frac{1.96}{\sqrt{n}}) = 0.216(1 + \frac{1.96}{\sqrt{6}}) = 0.389$

The condition to apply this formula is the sample size of at least 15. We need to show that despite our small sample size (6), the method still works pretty good. We can plot the data points as a scatter plot, alogside plot lines representing MLE, lower bound, and upper bound. We can see from the graph in Figure 3 that there is a chunk of data points in one place on the left side, which is supported by the highest points of the curves at the same point. However, it is hard to draw conclusions about how likely the data points are to come from the distributions because of the small sample size. We aso see that the lower bound curve does not go up much on the left side. Thus, the second plot represents likelihoods over different lambda values with corresponding confidence interval. We see that the right side of the interval does not acount for larger values of lambda and should be moved a little bit to the right. Besides that, the confidence interval is valid and appropriate.

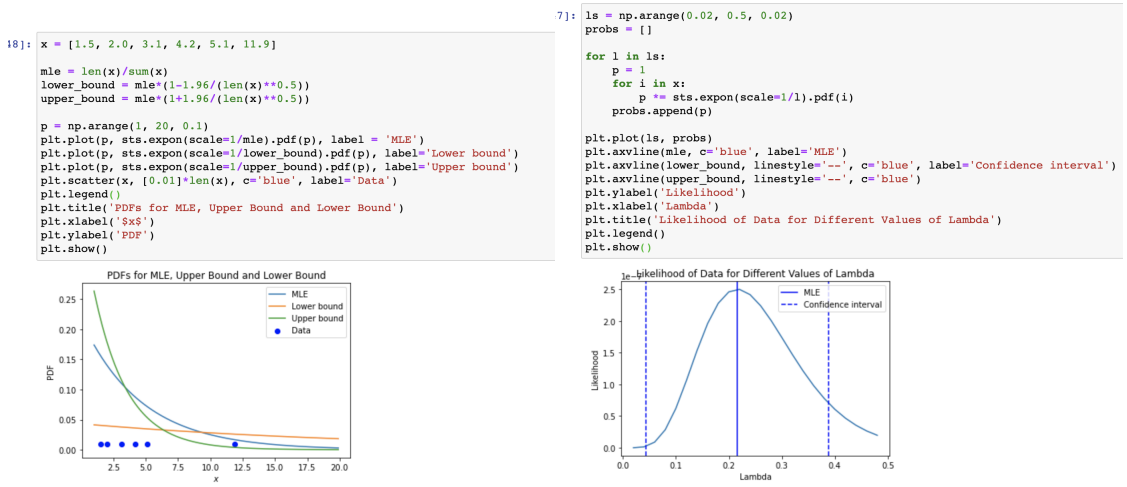

Figure 3: Justification of confidence interval

References

[1] Blitzstein, J. K., & Hwang, J. (2019). Introduction to probability. CRC Press/Taylor & Francis Group.

[2] Wikimedia Foundation. (2022, March 8). Exponential distribution. Wikipedia. Retrieved April 13, 2022, from https://en.wikipedia.org/wiki/Exponential_distribution

[3] Wikimedia Foundation. (2022, March 11). Continuous Uniform Distribution. Wikipedia. Retrieved April 15, 2022, from https://en.wikipedia.org/wiki/Continuous_uniform_distribution

[4] Wikimedia Foundation. (2022, March 30). Poisson distribution. Wikipedia. Retrieved April 15, 2022, from https://en.wikipedia.org/wiki/Poisson_distribution

[5] Wikimedia Foundation. (2021, November 1). Sum of normally distributed random variables. Wikipedia. Retrieved April 15, 2022, from https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables