Anmol Saraf
Email: 200070007@iitb.ac.in
Course: DS203 - Programming for Data Science Minor
Instructor: Dr. Manjesh K Hanawal

Assignment - 2
Roll No.: 200070007
Semester: Autumn 2021
Due Date: 22^{nd} August, 2021

Exercise 1
Let X and Y be independent exponential random variables with respective parameters λ_1 and λ_2. Find the distribution of the following.

- min(X, Y)
- max(X, Y)

Solution: For an exponential random variable X $\in [0, \infty)$, the PDF is given as-

$$f_X = \begin{cases} \lambda e^{-\lambda x} & if x \geq 0 \\ 0, & otherwise \end{cases}$$

Therefore, using this we calculate the Probability $P(X \leq x)$ as -

$$\begin{aligned} P(X \leq x) = \int_{-\infty}^{x} f_X &= \int_{-\infty}^{0} 0 + \int_{0}^{x} f_X \\ &= 0 + [-e^{-\lambda x}]_0^x \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Furthermore, $P(X \leq x) = 1 - P(X > x)$ can be obtained using the basic properties of probability.

Part 1 Let $Z = min(X, Y)$, now for $Z > x$, $min(X, Y) > x$ and for it to be each of X and Y has to be greater than x.

$$\begin{aligned} P(Z > x) &= P(min(X, Y) > x) \\ &= P(X > x \cap Y > x) \end{aligned}$$

Now, since X and Y are independent exponential random variables with parameters λ_1 and λ_2 respectively. We can write,

$$\begin{aligned} P(Z > x) &= P(X > x) \cdot P(Y > x) \\ &= e^{-\lambda_1 x} \cdot e^{-\lambda_2 x} \\ &= e^{-(\lambda_1 + \lambda_2)x} \end{aligned}$$

Since by definition, $F_X(x) = P(X \leq x) = 1 - P(X > x)$, the distribution of $min(X, Y)$ will be-

$$F_X(x) = 1 - e^{-(\lambda_1 + \lambda_2)x}$$

Part 2 Let $W = max(X, Y)$, now for $W \leq x$, $max(X, Y) \leq x$ and thus for this each of X and Y has to be smaller than or equal to x.

$$\begin{aligned} P(W \leq x) &= P(max(X, Y) \leq x) \\ &= P(X \leq x \cap Y \leq x) \end{aligned}$$

Similarly as Part 1, using the independence of X and Y,

$$\begin{aligned} P(W \leq x) &= P(X \leq x) \cdot P(Y \leq x) \\ &= (1 - e^{-\lambda_1 x}) \cdot (1 - e^{-\lambda_2 x}) \end{aligned}$$

Therefore, the distribution of $max(X, Y)$ will be-

$$F_X(x) = 1 - e^{-\lambda_1 x} - e^{-\lambda_2 x} + e^{-(\lambda_1+\lambda_2)x}$$

Exercise 2
A bag contains 3 white, 6 red and 5 blue balls. A ball is selected at random, it's color is noted and is then replaced in the bag before making the next selection. In all 6 selections are made. Let X = the number of white balls selected and Y = number of blue balls selected. Find E[X|Y = 3].

Solution: Since 3 of the 6 balls selected are blue, which lefts 3 balls which can be either white or red each. Therefore, the random variable $X = \{0, 1, 2, 3\}$. Since expectation $E(X)$, is defined as $E(X) = \sum_{x_i \in X} x_i P_X(x_i)$. So for $E[X|Y = 3]$,

$$E[X|Y = 3] = 0{\cdot}P(X = 0|Y = 3)+1{\cdot}P(X = 1|Y = 3)+2{\cdot}P(X = 2|Y = 3)+3{\cdot}P(X = 3|Y = 3)$$

Now we find all the terms needed to calculate the expectation. Firstly we calculate for any n number of white balls in terms of n.

We know that $P(blue) = \dfrac{5}{14}$, $P(white) = \dfrac{3}{14}$ and $P(red) = \dfrac{6}{14}$. Thus using these we calculate,

$$\begin{aligned}
P(X = n|Y = 3) &= \frac{P(X = n, Y = 3)}{P(Y = 3)} \\
&= \frac{{}^6C_3 {}^3C_n \left(\frac{3}{14}\right)^n \left(\frac{5}{14}\right)^3 \left(\frac{6}{14}\right)^{3-n}}{{}^6C_3 \left(\frac{5}{14}\right)^3 \left(\frac{9}{14}\right)^3} \\
&= \frac{{}^3C_n \cdot (2)^{3-n}}{(3)^3}
\end{aligned}$$

Therefore, using this we calculate expectation as-

$$\begin{aligned}
E[X|Y = 3] &= P(X = 1|Y = 3) + 2 \cdot P(X = 2|Y = 3) + 3 \cdot P(X = 3|Y = 3) \\
&= \frac{{}^3C_1(2)^2}{(3)^3} + 2 \cdot \frac{{}^3C_2(2)^1}{(3)^3} + 3 \cdot \frac{{}^3C_3(2)^0}{(3)^3} \\
&= \frac{4}{9} + \frac{4}{9} + \frac{1}{9} \\
&= 1
\end{aligned}$$

Exercise 3
If X_1 and X_2 are independent binomial random variables with respective parameters (n_1, p) and (n_2, p). Caluculate the conditional probability mass function of X_1 given that $X_1+X_2 = m$.

Solution: For a binomial random variable X, with respective parameters as (n, p) has its PMF as $P(X = i) = {}^{n}C_i p^i (1-p)^{n-i} \quad \text{for } 0 \le i \le n$. We have to find the conditional PMF of X_1 given that $X_1 + X_2 = m$, i.e. $P(X_1 = i | X_1 + X_2 = m) \quad \text{for } 0 \le i \le m$.

Thus to find this we first calculate $P(X_1 + X_2 = m)$

$$P(X_1 + X_2 = m) = \sum_{i=0}^{m} P(X_1 = i, X_2 = m - i)$$

Since X_1 and X_2 are independent,

$$\begin{aligned} P(X_1 + X_2 = m) &= \sum_{i=0}^{m} P(X_1 = i) \cdot P(X_2 = m - i) \\ &= \sum_{i=0}^{m} ({}^{n_1}C_i p^i (1-p)^{n_1 - i}) \cdot ({}^{n_2}C_{m-i} p^{m-i} (1-p)^{n_2 + i - m}) \\ &= p^m (1-p)^{n_1 + n_2 - m} \cdot \sum_{i=0}^{m} {}^{n_1}C_i \cdot {}^{n_2}C_{m-i} \end{aligned}$$

$\sum_{i=0}^{k} {}^{a}C_i \cdot {}^{b}C_{k-i} = {}^{a+b}C_k$ is given by Vandermonde's Identity. So,

$$P(X_1 + X_2 = m) = {}^{n_1+n_2}C_m \cdot p^m (1-p)^{n_1+n_2-m}$$

Which is a binomial RV with respective parameters as $(n_1 + n_2, p)$. Now, to calculate the conditional PMF of X_1,

$$\begin{aligned} P(X_1 = i | X_1 + X_2 = m) &= \frac{P(X_1 = i, X_1 + X_2 = m)}{P(X_1 + X_2 = m)} \\ &= \frac{P(X_1 = i, X_2 = m - i)}{P(X_1 + X_2 = m)} \\ &= \frac{P(X_1 = i) \cdot P(X_2 = m - i)}{P(X_1 + X_2 = m)} \\ &= \frac{({}^{n_1}C_i \cdot p^i (1-p)^{n_1 - i}) \cdot ({}^{n_2}C_{m-i} \cdot p^{m-i} (1-p)^{n_2 + i - m})}{{}^{n_1+n_2}C_m \cdot p^m (1-p)^{n_1+n_2-m}} \\ &= \frac{{}^{n_1}C_i \cdot {}^{n_2}C_{m-i}}{{}^{n_1+n_2}C_m} \end{aligned}$$

Exercise 4
Give an example of two random variables X and Y that are uncorrelated but not independent.

Solution: Let A and B be two independent discrete random variables each taking values in $\{0, 1\}$ such that,

$$P(A = 0) = P(A = 1) = \frac{1}{2}$$

$$P(B = 0) = P(B = 1) = \frac{1}{2}$$

Let $X = A + B$ and $Y = A - B$ be two random variables.
Thus,

$$\begin{aligned}
E(X) &= \sum_{i=0}^{2} x_i P_X(x_i) \\
&= 0 \cdot P(A+B=0) + 1 \cdot P(A+B=1) + 2 \cdot P(A+B=2) \\
&= 1 \cdot \left(\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}\right) + 2 \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right) \\
&= 1
\end{aligned}$$

$$\begin{aligned}
E(Y) &= \sum_{i=-1}^{1} y_i P_Y(y_i) \\
&= (-1) \cdot P(A-B=-1) + 0 \cdot P(A-B=0) + 1 \cdot P(A-B=1) \\
&= (-1) \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
E(XY) &= \sum_{i=-1}^{1} xy_i P_{XY}(xy_i) \\
&= (-1) \cdot P(A^2-B^2=-1) + 0 \cdot P(A^2-B^2=0) + 1 \cdot P(A^2-B^2=1) \\
&= (-1) \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) \\
&= 0
\end{aligned}$$

Therefore, the covariance of X and Y, $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$. Which means that X and Y are uncorrelated.
Now for checking if they are independent or not, we find $P(X = 0, Y = 0)$, which has the only case of $A = 0$ and $B = 0$. Thus $P(X = 0, Y = 0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. But,

$$P(X=0) = P(A+B=0) = P(A=0, B=0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$\begin{aligned}
P(Y=0) = P(A-B=0) &= P(A=0, B=0) + P(A=1, B=1) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{2}
\end{aligned}$$

Therefore, $P(X = 0)P(Y = 0) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} \neq P(X = 0, Y = 0)$.
Thus, this is an example of two random variables which are uncorrelated but are not independent.

Exercise 5
Suppose X is a Poisson random variable with mean λ. The parameter λ is itself a random variable whose distribution is exponential with mean 1. Show that P{X = n} = $(1/2)^{n+1}$.

Solution: Since X is a Poisson random variable with mean λ. In Poisson distribution the mean of the distribution is equal to the parameter of the distribution. So, λ is also the parameter of the distribution.

Also, λ is in itself an exponential random variable with mean 1. In Exponential distribution the mean of the distribution is equal to $\dfrac{1}{parameter}$ of the distribution. So, its parameter will be 1.

So, $P(X = n) = \dfrac{e^{-\lambda}\lambda^n}{n!}$ and λ will have a PDF as $f_\lambda(x) = 1 \cdot e^{-1 \cdot x} = e^{-x}$ for $x \geq 0$.

Now, using the Total Probability Law,

$$\begin{aligned} P(X = n) &= \int_0^\infty P(X = n|\lambda = x) f_\lambda(x) dx \\ &= \int_0^\infty \left(\frac{e^{-x}x^n}{n!}\right) \cdot e^{-x} dx \\ &= \frac{1}{n!} \cdot \int_0^\infty x^n e^{-2x} dx \end{aligned}$$

Substituting $2x$ as t in the integral.

$$\begin{aligned} P(X = n) &= \frac{1}{(2)^{n+1} n!} \cdot \int_0^\infty t^n e^{-t} dt \\ &= \frac{1}{(2)^{n+1} n!} \cdot \Gamma(n+1) \end{aligned}$$

Here $\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt$ by definition and for integer n, $\Gamma(n) = (n-1)!$.

Therefore,

$$P(X = n) = \left(\frac{1}{2}\right)^{n+1}$$

Exercise 6
Suppose X and Y have joint density function $f_{X,Y}(x, y) = c(1 + xy)$ if $2 \leq x \leq 3$ and $1 \leq x \leq 2$, and $f_{X,Y}(x, y) = 0$ otherwise.

1. Find c.
2. Find f_X and f_Y.

Solution: The joint density function of X and Y is given as,

$$f_{X,Y}(x, y) = \begin{cases} c(1 + xy), & if\ 2 \leq x \leq 3\ and\ 1 \leq x \leq 2 \\ 0, & otherwise \end{cases}$$

Now, to find c, we know that the double integral of $f_{X,Y}(x, y)$ should be equal to 1 as to fulfill the normalization property of Probability. So,

$$\begin{aligned}
1 &= \int_{\mathbb{R}^2} f_{X,Y}(x,y) \\
&= \int_1^2 \int_2^3 f_{X,Y}(x,y) \cdot dx \cdot dy \\
&= \int_1^2 \int_2^3 c(1+xy) \cdot dx \cdot dy \\
&= \int_1^2 \left[cx + \frac{cy}{2}x^2\right]_2^3 \cdot dy \\
&= \int_1^2 \left(c + \frac{5cy}{2}\right) \cdot dy \\
&= \left[cy\frac{5c}{4}y^2\right]_1^2 \\
&= c + \frac{15c}{4} = \frac{19c}{4} \\
\implies c &= \frac{4}{19}
\end{aligned}$$

For finding f_X and f_Y we integrate $f_{X,Y}$ with respect to y and x respectively.

$$\begin{aligned}
f_X &= \int_1^2 f_{X,Y}(x,y) \cdot dy \\
&= \int_1^2 \frac{4}{19}(1+xy) \cdot dy \\
&= \left[\frac{4}{19}\left(y + \frac{x}{2}y^2\right)\right]_1^2 \\
&= \frac{4}{19} \cdot \left(1 + \frac{3x}{2}\right) \\
&= \frac{4+6x}{19}
\end{aligned}$$

Similarly for f_Y,

$$\begin{aligned}
f_Y &= \int_2^3 f_{X,Y}(x,y) \cdot dx \\
&= \int_2^3 \frac{4}{19}(1+xy) \cdot dx \\
&= \left[\frac{4}{19}\left(x + \frac{y}{2}x^2\right)\right]_2^3 \\
&= \frac{4}{19} \cdot \left(1 + \frac{5y}{2}\right) \\
&= \frac{4+10y}{19}
\end{aligned}$$

Therefore,

$$c = \frac{4}{19}, \quad f_X(x) = \frac{4+6x}{19}, \quad f_Y(y) = \frac{4+10y}{19}$$

Exercise 7
An insurance company supposes that the number of accidents that each of its policyholders will have in a year is Poisson distributed, with the mean of the Poisson depending on the policyholder. If the Poisson mean of a randomly chosen policyholder has a gamma distribution with density function,

$$g(\lambda) = \lambda e^{-\lambda}, \qquad \lambda \geq 0$$

What is the probability that a randomly chosen policyholder has exactly n accidents next year?

Solution: Let X be a Poisson distributed random variable whose mean is a random variable with gamma distribution.
Let ϕ be the mean of X, as we know the mean of a Poisson RV is equal to the parameter of the RV.
Therefore, $P(X = i) = \dfrac{e^{-\phi}\phi^i}{i!}$ where ϕ has a PDF of $g(\lambda) = \lambda e^{-\lambda}$.
Thus, using the Total Probability Law,

$$\begin{aligned} P(X = n) &= \int_0^\infty P(X = n|\phi = \lambda) \cdot g(\lambda) \cdot d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda}\lambda^n}{n!} \cdot \lambda e^{-\lambda} \cdot d\lambda \\ &= \frac{1}{n!} \cdot \int_0^\infty \lambda^{n+1} e^{-2\lambda} \cdot d\lambda \end{aligned}$$

Substituting 2λ as t in the integral,

$$\begin{aligned} P(X = n) &= \frac{1}{(2)^{n+2} \cdot n!} \cdot \int_0^\infty t^{n+1} e^{-t} \cdot dt \\ &= \frac{1}{(2)^{n+2} \cdot n!} \cdot \Gamma(n+2) \end{aligned}$$

Here $\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt$ by definition and for integer n, $\Gamma(n) = (n-1)!$.

Therefore,

$$\begin{aligned} P(X = n) &= \frac{1}{(2)^{n+2} \cdot n!} \cdot (n+1)! \\ &= \frac{n+1}{2^{n+2}} \end{aligned}$$

Exercise 8
Suppose that the number of people who visit a yoga studio each day is a Poisson random variable with mean λ. Suppose further that each person who visits is, independently, female with probability p or male with probability 1 - p. Find the joint probability that exactly n women and m men visit the academy today.

Solution: Let N and M be random variables representing the number of women and men that visit the academy on a day. Now, let T be the total number of people visiting the academy. Thus, $T = N + M$, and T is a Poisson distributed random variable as given in the question with mean λ.

Therefore, $P(T=k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$, as the mean of Poisson distribution is the same as its parameter.

Now, using the Conditional Probability theorem, i.e., $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, we need to find $P(N=n, M=m)$ and we know $P(T=n+m)$ so

$$\begin{aligned} P(N=n, M=m) &= P(N=n, M=m|T=n+m)P(T=n+m) \\ &= P(N=n, M=m|T=n+m) \cdot \frac{e^{-\lambda}\lambda^{n+m}}{(n+m)!} \end{aligned}$$

Now to find $P(N=n, M=m|T=n+m)$, this is simply the probability of the total number of people of which n being women. Thus, this will be Binomial distributed in random variable N.

$$P(N=n, M=m, T=n+m) = {}^{n+m}C_n \cdot p^n \cdot (1-p)^m$$

Finally, we can write the probability $P(N=n, M=m)$ as,

$$\begin{aligned} P(N=n, M=m) &= P(N=n, M=m, T=n+m) \cdot \frac{e^{-\lambda}\lambda^{n+m}}{(n+m)!} \\ &= {}^{n+m}C_n \cdot p^n \cdot (1-p)^m \cdot \frac{e^{-\lambda}\lambda^{n+m}}{(n+m)!} \end{aligned}$$

Exercise 9
Let X_1, X_2, X_3 are RVs and a, b, c, d are constants. Show that

- $Cov(aX_1 + b, cX_2 + b) = acCov(X_1, X_2)$
- $Cov(X_1 + X_2, X_3) = Cov(X_1, X_3) + Cov(X_2, X_3)$

Solution: We know that the covariance of two random vairables A and B is given as, $Cov(A, B) = E(AB) - E(A)E(B)$.

Part 1
$Cov(aX_1 + b, cX_2 + b)$
$\implies E((aX_1 + b)(cX_2 + b)) - E(aX_1 + b)E(cX_2 + b)$
$\implies E(acX_1X_2 + abX_1 + bcX_2 + b^2) - [(aE(X_1) + b)(cX_2 + b)]$
$\implies acE(X_1X_2) + abE(X_1) + bcE(X_2) + b^2 - [acE(X_1)E(X_2) + abE(X_1) + bcE(X_2) + b^2]$
$\implies acE(X_1X_2) - acE(X_1)E(X_2)$
$\implies ac(E(X_1X_2) - E(X_1)E(X_2))$
$\implies ac \cdot Cov(X_1, X_2)$

Part 2
$Cov(X_1 + X_2, X_3)$
$\implies E((X_1 + X_2)X_3) - E(X_1 + X_2)E(X_3)$
$\implies E(X_1X_3 + X_2X_3) - (E(X_1) + E(X_2))E(X_3)$
$\implies E(X_1X_3) + E(X_2X_3) - E(X_1)E(X_3) - E(X_2)E(X_3)$
$\implies [E(X_1X_3) - E(X_1)E(X_3)] + [E(X_2X_3) - E(X_2)E(X_3)]$
$\implies Cov(X_1, X_3) + Cov(X_2, X_3)$

Exercise 10
You are given n = 100 i.i.d. samples generated from a random experiment. Let the estimate of mean from these samples is $\hat{\mu} = 0.45$. We know that true mean lies somewhere around $\hat{\mu}$ and we would like to find an interval (around $\hat{\mu}$) such that the true value lies in the interval with probability at least 0.95.

- What would be your (confidence) interval? Specify the method you used to come up with the interval.
- If you want the your confidence interval to shrink by half, how many more samples would you need? (the estimate could be different now)

Solution: Using the Confidence Interval formula, $P(|\hat{\mu} - \mu| > \epsilon) \leq 2e^{-n\epsilon^2}$. Where, $\hat{\mu} = 0.45$ is given. μ is $E(X_1)$ and its range is $(\hat{\mu} - \epsilon, \hat{\mu} + \epsilon)$. Also it is given that the value lies in the range with a probability of atleast 0.95. Which implies, $P(|\hat{\mu} - \mu| \leq \epsilon) \geq 0.95$.

Therefore, $P(|\hat{\mu} - \mu| > \epsilon) \leq 0.05$ will be the probability condition for it to not lie in its range. **Part 1** Thus, this shows that,

$$\begin{aligned}
2e^{-n\epsilon^2} &= 0.05 \\
e^{-n\epsilon^2} &= 0.025 \\
-n\epsilon^2 &= \ln(0.025) \\
-n\epsilon^2 &= -3.689 \\
\epsilon^2 &= \frac{3.689}{100} \\
\epsilon &= \sqrt{0.03689} = 0.192
\end{aligned}$$

Therefore, the confidence interval defined as $(\hat{\mu} - \epsilon, \hat{\mu} + \epsilon)$ will be $(0.258, 0.642)$.

Part 2 To shrink the confidence interval by half, i.e., $\epsilon' = \epsilon/2$, the total number of samples n' needed will change. As the probability of the true value lying in the interval does not change we can say,

$$2e^{-n\epsilon^2} = 2e^{-n'\epsilon'^2}$$

Thus, dividng by 2 and taking ln on both sides gives us,

$$n\epsilon^2 = n'\epsilon'^2$$

And since $\epsilon' = \epsilon/2$

$$n' = \frac{n\epsilon^2}{(\epsilon/2)^2} = 4n = 400$$

Thus we would need 300 more samples than before, or 400 samples in total to shrink the confidence interval by half.