Final exam

Farnoosh Koleini

2022-12-10

Instructions

Honor Code: This is a take-home exam. You may consult the textbook, lecture slides, homework solutions, and midterm exam solutions. Consulting with the students in the class or others is prohibited. What you submit should your own work. Please show all solution steps. You will not receive any credit if you just submit the final answer without the solution steps.

If your solution requires writing R code, please comment your code.

Use this document to enter your solutions. Please type your response to a question right below the question text. Compile this document to generate an HTML or PDF output. Upload your HTML/PDF document to MS Teams. In the preamble above, change `Your Name` to your name.

Questions

Counting/ Permutations/ Combinations

A jar of 40 balls contains 10 red color balls, 10 green color balls, 10 blue color balls, and 10 purple color balls. If you draw four balls from the jar randomly **without replacement**, what is the probability that all four colors are in the draw? Write a Monte Carlo simulation to support your analytic result.

Solution: analytic solution goes here.

This problem is quite simple, we have 40 balls and we need to draw 4 balls. So the sample space is choosing 4 balls from 40 balls $ {40 } = 91390$. Note: all of these four picking balls are independent of each other and each of them is $ {10 } = 10$ he answer would be:

$$P(X) = \frac{10\times10\times10\times10}{\binom{40}{4}} = 0.109396$$

And the simulation result shows that, it would be around 0.10. If we icrease the number of permutations we will get a more accurate answer.

A Monte Carlo simulation to provide support for the analytical result:

```
# unique() reduces a vector to its unique elements (removes duplicates)

nTrials <- 1000000
jar <- (c(rep(1, 10), rep(2, 10), rep(3, 10), rep(4, 10)))
# jar
success <- 0
for (i in 1:nTrials){
  aDraw <- sample(jar, 4, replace = FALSE)
  if (length(unique(aDraw)) == 4){
    success <- success + 1
    # print(aDraw)
  }
}
cat(paste0("Probability that all four colors are in the draw: \f\n", success/nTrials))
```

```
## Probability that all four colors are in the draw: 
## 0.109241
```

■

Conditional probability and independence

Suppose that in a certain country 10% of the elderly people have diabetes. It is also known that 30% of the elderly people are living below the poverty level, and 35% of the elderly population falls into at least one of these categories.

a. Given that a randomly selected elderly person is living below the poverty level, what is the probability that she/he has diabetes?

Solution:

The key note here is the conditional probability which is $P(A|B) = \frac{P(A\cap B)}{P(B)}$. Now it is time to mention what is event A and B. Let's say event A is a person has a diabetes and event B is defined as randomly selected elderly person is living below the poverty level. So based on the information provided in the question we just need to find the conditional probability of A and B.

$P(A) = 0.1$, and $P(B) = 0.3$ and because $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.35$. So the probability of $P(A \cap B) = 0.40 - 0.35 = 0.05$.

So $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.35} = 0.142$

b. Are the events “has diabetes” and “living below the poverty level” disjoint in this elderly population? Explain.

Solution: Disjoint is different from independent. Disjoint means if probability of A intersection B is NULLSET. So if the answer for intersection of A and B is not zero means that these tow A and B are not disjoint.

c. Are the events “has diabetes” and “living below the poverty level” independent in this elderly population? Explain.

Solution:

If we calculate the conditional probability which is done in part A. If $P(A|B) = P(A)$ this sentence is true and they are independent but the calculation shows these two are not equal, so these A and B are not independent.

■

Discrete random variables and expected values

A quiz consists of two multiple choice questions with choices (i), (ii), and (iii) for each. If an unprepared student marks answers at random, what are the probabilities of the following events?

a. Both answers are correct.

Solution:

If we find the sample space $\Omega = S = CC, CW, WC, WW$, Now we can draw a tree and answer this question easily! $P(C) = \frac{1}{3}$, and $P(W) = \frac{2}{3}$.

Both answer correct would be CC that its probability would be $\frac{1}{9}$.

b. Exactly one correct answer.

Solution:

So CW and WC are both correct for this situation. The final answer would be $2 \times \frac{1}{9} = \frac{2}{9}$

c. Both answers wrong.

Both answer correct would be WW that its probability would be $\frac{4}{9}$.

Solution:

d. Let Y be the random variable representing the number of correct answers on the quiz. What is $\mathbb{E}(Y)$?

Solution:

It is weighted average:

$\mathbb{E}(Y) = P_0V_0 + P_1V_1 + P_2V_2 = 0 + \frac{1}{3} \times 1 + \frac{1}{3} \times 2 = 1$

■

Continuous random variable/probability density function

The random variable Y has probability density

$$\mathbb{f}(y) = cye^{-y}, \quad y > 0$$

(a) What is c?

Solution:

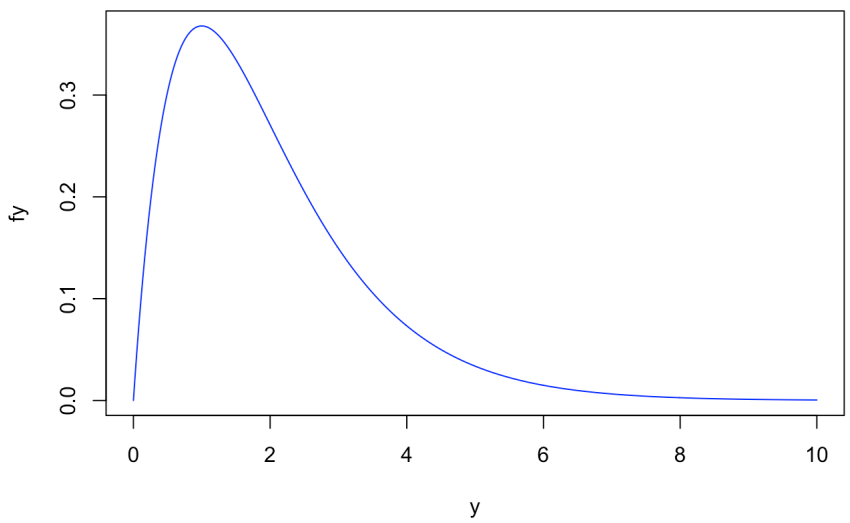
$\int cye^{-y}dy = 1$ where y>0. We can solve this integration easily in Wolfram alpha website or just solving this by ourselves.
$\int cye^{-y}dy = -ce^{-y}(y+1)$ if we calculate this for y = 0. It would be $-c \times 1 \times 1 = 1$ So $C = 1$.

b. Sketch the density.

Solution:

We can write a R code for example:

```
y <- seq(0,10, by = 0.01)
fy <- 1* y* exp(-y)
plot (y, fy, type = "l", col = "blue")
```



c. What is $\mathbb{E}(Y)$? (Hint: Recall the gamma function)

Solution:

$cye^{-y}dy = 1ye^{-y}dy $ This pattern can be easily being seen without lots of calculations. We need to look at the sketch to see where the balance is. The balance is about 1.6. This value is the expected value. And the point here is the mass right side and left side of the falcon is the same and the balance point is obvious in the plot.

d. What is the mode of the density (i.e., the value of y for which $f(y)$ is largest)?

Solution:

One way to do this is you know the function, first we take the first derivative of our function and equal to 0. Then solving for y. Because our function is the function of y we take the first derivative and equal to 0 then solving for y. So when you find the y value, it is going to be the highest point in the graph. It is the mode. $\frac{d(ye^{-y})}{dy} = -e^{-y}(y-1)$ If we calculate this for y = 0. It is going to be 1 and this is the mode.

■

Finding Median using Cumulative Distribution Function

Let Y be a continuous random variable with PDF

$$\mathbb{f}_Y(y) = y\, e^{-y^2/2}, \quad y > 0$$

(a) Find the CDF for Y.

Solution:

We have the pdf in the question, now it is time to find the CDF. We need to do integration of pdf. $\int y\, e^{-y^2/2}$ To calculate this we need a dummy variable. $\int_0^y w\, e^{-w^2/2}$ We can use wolfram alpha or we can use another suitable variable like $u = \frac{w^2}{2}, du = wdw$. So the answer is:

$\int_0^y w\, e^{-w^2/2} = -e^{\frac{-w^2}{2}}(w^2-1)$ for w=0 would be 1 and for w = y would be $-e^{\frac{-y^2}{2}}(y^2-1)$ and the subtraction would be $-e^{\frac{-y^2}{2}}(y^2-1)-1$

b. Find the median (that is, the 50th percentile) of the distribution.

Solution:

$F_Y(y_1) = 0.5$ and if we solve this for y, it gives us the median. $-e^{\frac{-y^2}{2}}(y^2-1)-1 = 0.5, -e^{\frac{-y^2}{2}}(y^2-1) = 1.5$ The value of y from this formula would be the median or the 50th percentile.

■

Normal Random Variable

Filaments made at a factory are expected to contain 2.75mg of *chromelite*. The amount of chromelite in a filament is actually a random variable due to the randomness in the manufacturing process. If a filament has more than 2.77mg or less than 2.73mg of chromelite, the filament must be discarded. Machine A makes filaments with a chromelite distribution which is normally distributed with mean $\mu_A = 2.75\text{mg}$ and standard deviation $\sigma_A = 0.01\text{mg}$. Machine B makes filaments with a chromelite distribution which is normally distributed with mean $\mu_A = 2.76\text{mg}$ and standard deviation $\sigma_A = 0.005\text{mg}$. Which machine is better, in terms of a smaller proportion of discarded filaments?

Solution:

This question is kind of simple which means it does not need many calculations. We are given two normal distributions. And we are given enough information to decide machine A versus machine B, which one is performing better in terms of smaller proportion of discarded filaments. if we create two diagrams for both A and B, the first diagram is A and the second would be B. Based on the graphs and the sigma values provided in the question we can use a hypothesis testing. Machine B has better performance because the mean value is greater and the sigma which is related to variance is smaller. So Machine B is better in terms of a smaller proportion of discarded filaments.

■

Bivariate Continuous Joint Distribution

Let X_1 and X_2 be jointly continuous random variables with PDF

$$\mathbb{f}\,(x_1, x_2) = c\, x_1^2\, x_2^2, \quad 0 < x_1 + x_2 < 1,\ x_1 > 0,\ x_2 > 0$$

(a) Find c

Solution:

The area is equal to 1 and $c \times \int_0^1 \int_0^{1-y_2} y_1^2 y_2^2 \; dy_2 \; dy_1$, first based on the y2 we need to do integration.

$c \times \int_0^1 \int_0^{1-y_2} y_1^2 y_2^2 \; dy_2 \; dy_1 = \int_0^1 2(1-y_2)^3 y_1^2 \; dy_1 = 2(1-y_2)^3 \int_0^1 y_1^2 \; dy_1 = 2(1-y_2)^3 \times 2 = 4c(1-y_2)^3 = 1$

So $c = \frac{1}{4(1-y2)^3}$

b. Find $\mathbb{P}\,(X_2 > 1/2)$

Solution:

We need to know c to solve this problem which is already calculated in the section (a). If you think about x1 = x2, and create a triangle of that which has three corners, (0,0), (0,x2), and (x1,0). The right region of that triangle would be x1>x2.

$\frac{1}{4(1-y2)^3} \int_{y_2}^1 \int_0^{1-y2} y_1^2 y_2^2 \; dy_2 \; dy_1 = \frac{1}{4(1-y2)^3} \int_{y_2}^1 -2(1-y_2)^3 = \frac{-2(1-y_2)^3}{4(1-y2)^3}(2y_2^3 - 2) = 1 - y_2^3$

c. Find $\mathbb{P}\,(X_1 > X_2)$

Solution:

Again here we have a double integration.

$c \times \int_{1/2}^1 \int_{y_2}^{1-y_2} y_1^2 y_2^2 \; dy_2 \; dy_1 = c \int_{1/2}^1 y_1^2 (2y_2^3 - 2(1-y_2)^3) \, dy_1 = \frac{1}{4(1-y2)^3} \times (2y_2^3 - 2(1-y_2)^3) \int_{1/2}^1 y_1^2 \; dy_1 = \frac{1}{4(1-y2)^3} \times (2y_2^3 - 2(1-y_2)^3) \times (2/8 - 2) = \frac{}{4(1-}$

d. Find $\mathbb{P}\,(X_2 > X_1)$.

Solution:

The answer for part d is 1 - answer of the part c. So now we know the answer of the part c which is $\frac{1}{4(1-y2)^3} \times (2y_2^3 - 2(1-y_2)^3) \times (-14/8)$.

So answer for part d is going to be $ 1 - (2y_2^3 - 2(1-y_2)^3)(-14/8)$

e. Provide support for analytic solutions to (b)-(d) using Monte Carlo simulations.

Solution:

What we need here is $0 < x_1 + x_2 < 1$. For each pairs of x1 and x2 we have 1.

```
x1 <- runif(1000000)
x2 <- runif(1000000)
```

■

Transformation of a Random Variable using MGF Method

Let $U \sim \text{unif}(0, 1)$ and $Y = -\log(U)/5$.

a. Using the method of moment generating functions, find the PDF of Y.

Solution:

We have two random variables Y and U. Y is a function of the first random variable, U. First thing is what is the moment generating function, $E[e^{ty}] = E[e^{-tlog(U)/5}] = \int_0^1 e^{-tlog(U)/5} du = \int_0^1 e^{+log(u^{-t/5})} = \int_0^1 u^{-t/5} du = \frac{1}{1-t/5} = \frac{5}{5-t}$

b. Generalize your derivation to $Y = -\log(U)/a$.

Solution:

Instead of 5 we have a there, and we can actually say that $\frac{a}{a-t}$.

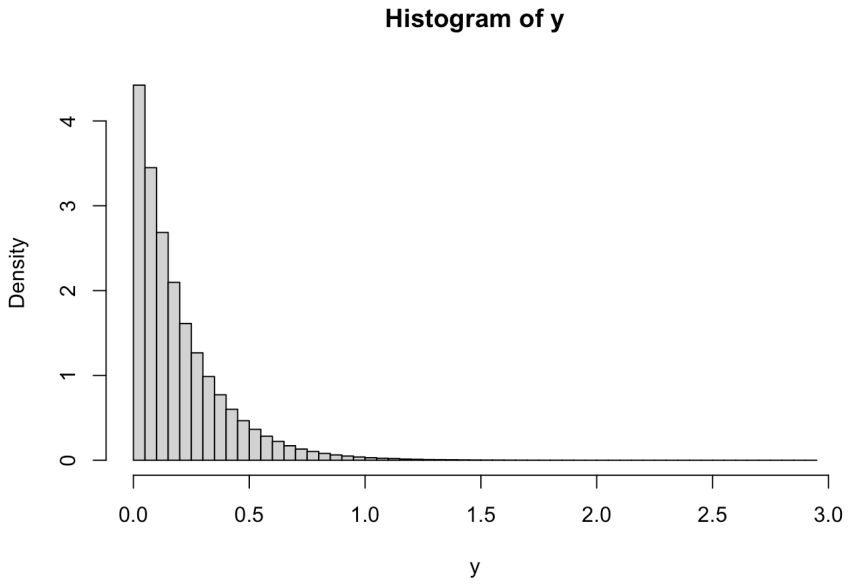
c. Provide support for your analytic solution to (a) using a Monte Carlo simulation.

Solution:

We can generate:

```
u <- runif(1000000)
y <- -log (u)/5

hist(y,freq = FALSE, breaks = 50)
```



■

Functions of Random Variables (CDF Technique)

Y is a random variable with PDF

$$\mathbb{f}_\theta(y) = \theta(1-y)^{\theta-1}, \quad y \in (0,1), \theta > 0$$

Let $W = -\log(1-Y)$ and find the PDF of W.

Solution:

We have random variable y, it belongs to 0 and 1. The density is given by $\mathbb{f}_\theta(y) = \theta(1-y)^{\theta-1}$. So we need to find the pdf of the w here. We are going to use the cumulative distribution function. By definition
$F_W(w) = P(W \leq w) = P(-log(1-Y) \leq w) = P(Y \leq e^w + 1) = P(Y \leq 1 - e^{-w})$

We are evaluating the cumulative distribution function of Y at $1 - e^{-w}$. So therefore this is going to be $F_Y(1 - e^{-w})$. What we have here is we have the CDF of w, we need the PDF. So we need to differentiate this CDF. So if we do that, what happens is the cumulative distribution function is $f_y(1 - e^{-w})$ after simplification, the thing we have is $f_W(w) = \theta e^{-w\theta}$.

■

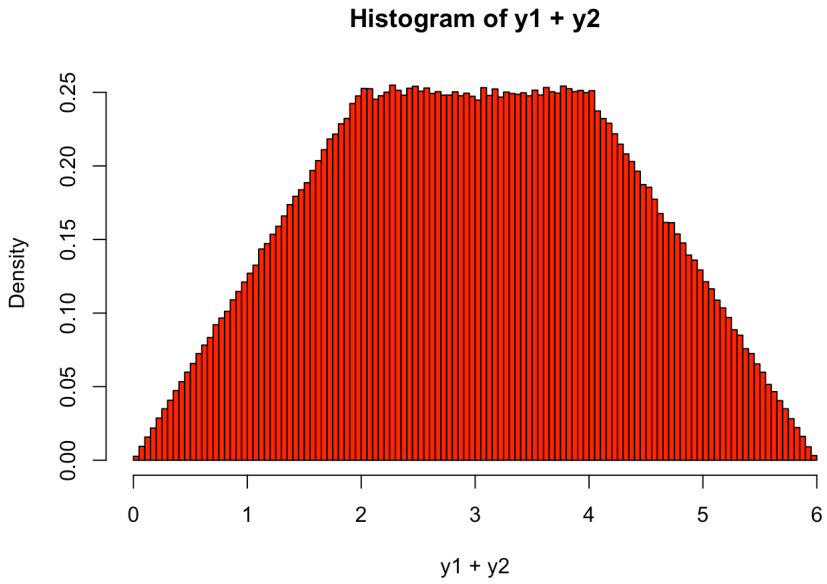
Transformations of Jointly Distributed Random Variables

The random variables X_1 and X_2 have joint density that is uniform on the area bounded by the lines $x_2 > 0, x_1 + x_2 < 4$, and $x_2 < x_1$. Find the PDF of $U = X_1 + X_2$ and sketch it. Provide support for your analytic solution using a Monte Carlo simulation.

Solution:

We have two random variables here, X1 and X2. What we can do is, finding the pdf the new random variable U which is the function of X1 and X2. This question is quite similar to x1+x2 < 1. We use the cumulative distribution function technique here to find $U = X_1 + X_2, f_U(u) = \frac{u}{8}$ where $u = (0, 4)$. Now we need to check is using monte carlo simulation.

```
y1 <- runif(1000000,0,4)
y2 <- runif(1000000,0,2)
index <- y1 + y2 < 4 & y1 > y2
newy1 <-y1[index]
newy2 <-y2[index]
hist(y1+y2, freq = FALSE, breaks = 100, col = "red")
```



■