STAT5000_HW4

October 16, 2023

1 Homework #4

See Canvas for HW #4 assignment due date. Complete all of the following problems. Ideally, the theoretical problems should be answered in a Markdown cell directly underneath the question. If you don't know LaTex/Markdown, you may submit separate handwritten solutions to the theoretical problems, but please see the class scanning policy in the syllabus. Please do not turn in messy work. Computational problems should be completed in this notebook (using the `R` kernel). Computational questions may require code, plots, analysis, interpretation, etc. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

1.1 A. Theoretical Problems

1.2 A.1 The median of the exponential distribution

- In class, we defined the mean of a random variable X with probability density function $f(x)$ to be $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.
- We also defined the median of a random variable X with probability density function $f(x)$ to be the value m such that $P(X \leq m) = 0.5$.

The exponential distribution has a density function

$$f(x) = \lambda e^{-\lambda x},$$

for $x \geq 0$, where $\lambda > 0$ is the rate parameter. We showed in class that if $X \sim Exp(\lambda)$, then $E(X) = \frac{1}{\lambda}$.

A.1(a) [5 points] Let $\lambda = 1$. Prove that the median of $X \sim Exp(1)$ is $\log(2)$ (on this problem, the notation $\log(\cdot)$ denotes the natural logarithm).

$$\int_0^m e^{-x}\, dx = 0.5$$

$$1 - e^{-m} = 0.5$$

$$e^{-m} = 0.5$$

$$\ln(e^{-m}) = \ln(0.5)$$

$$-m = \ln(0.5)$$

$$m = -\ln(0.5)$$

$$m = \ln(2)$$

A.1(b) [5 points] Let $\lambda = 2$. Prove that the median of $X \sim Exp(2)$ is $\frac{\log(2)}{2}$.

$$\int_0^m 2e^{-2x}\, dx = 0.5$$

$$1 - e^{-2m} = 0.5$$

$$e^{-2m} = 0.5$$

$$\ln(e^{-2m}) = \ln(0.5)$$

$$-2m = \ln(0.5)$$

$$m = \frac{\ln(0.5)}{-2}$$

$$m = \frac{\ln(2)}{2}$$

A.1(c) [7 points] Prove that for $X \sim Exp(\lambda)$, the median is given by $\frac{\log(2)}{\lambda}$

$$\int_0^m \lambda e^{-\lambda x}\, dx = 0.5$$
$$-\frac{1}{\lambda} e^{-\lambda x}\Big|_0^m = 0.5$$
$$-\frac{1}{\lambda}(e^{-\lambda m} - 1) = 0.5$$
$$1 - e^{-\lambda m} = 0.5$$
$$e^{-\lambda m} = 1 - 0.5 = 0.5$$
$$\ln(e^{-\lambda m}) = \ln(0.5)$$
$$-\lambda m = \ln(0.5)$$
$$m = \frac{\ln(0.5)}{-\lambda}$$
$$m = \frac{\ln(2)}{\lambda}$$

1.3 A.2 [8 points] Expected values and bets

Let $X =$ the outcome when a fair die is rolled once. Suppose that, before the die is rolled, you are offered a choice:

Option 1: a guarantee of $1/4$ dollars (whatever the outcome of the roll)

Option 2: $h(X) = 1/X$ dollars. Which option would you prefer?

Justify your answer.

For option 2, the expected payout per roll can be calculated using:

$$E(X_2) = p * (\frac{1}{6} + \frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2} + \frac{1}{1})$$

p is the probability of rolling each face on a fair die, which is 1/6.

$$E(X_2) = \frac{49}{120} = 0.408333$$

Since $0.408333 > 0.25$, option 2 is the better choice.

1.4 A.3 [8 points] More fraud

Recall the random variable from the previous homework: X is the leading digit of a randomly selected number from a large accounting ledger. The PMF was defined by:

$$P(X = x) = f(x) = \log_{10}\left(\frac{x+1}{x}\right), \quad x = 1, 2, ..., 9.$$

Give an expression for $E[X]$, and then calculate it.

$$E[X] = \sum_{x=1}^{9} x \cdot P(X = x)$$

$$E[X] = \sum_{x=1}^{9} x \cdot \log_{10}\left(\frac{x}{x+1}\right)$$

[ ]:

```
pmf <- function(x) {
  return(log10((x + 1) / x))
}

expected_value <- sum(1:9 * pmf(1:9))
expected_value
```

```
3.44023696712321
```

1.5 A.4 Mixture distributions

Most of the distributions we've looked at so far have been unimodal, flat, or monotonically increasing/decreasing. Let's take a look at a class of distributions that are multimodal.

Let $X_1, ..., X_n$ be continuous random variables with PDFs $f_{X_i}(x)$, $i = 1, ..., n$. Define a *mixture distribution* to be a random variable X with pdf

$$f_X(x) = \sum_{i=1}^{n} \alpha_i f_{X_i}(x)$$

where α_i are nonnegative real numbers such that $\sum_{i=1}^{n} \alpha_i = 1$. You can think of $f_X(x)$ as being a weighted average of the PDFs $f_{X_i}(x)$.

A.4(a) [4 points] Show that $f_X(x)$ is a PDF.

$$1. \quad f_X(x) \geq 0 \text{ for all } x. \tag{1}$$

2. To show that the integral of $f_X(x)$ over the entire range of x is equal to 1, we need to calculate: (2)

$$\int_{-\infty}^{\infty} f_X(x)\, dx. \tag{3}$$

Subsituting f_X:

$$\int_{-\infty}^{\infty} \sum_{i=1}^{n} \alpha_i f_{X_i}(x)\, dx. \tag{4}$$

And:

$$\sum_{i=1}^{n} \alpha_i \int_{-\infty}^{\infty} f_{X_i}(x)\, dx. \tag{5}$$

Since f_(X_i) are PDFs, each term of the integral above equals 1. Therefore, the sum of these terms is n:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = \sum_{i=1}^{n} \alpha_i \int_{-\infty}^{\infty} f_{X_i}(x)\, dx = \sum_{i=1}^{n} \alpha_i \cdot 1 = \sum_{i=1}^{n} \alpha_i = 1. \tag{6}$$

Therefore, f_X(x) satisfies both conditions and is a PDF.

A.4(b) [4 points] Find the expected value of X.

The expected value is a linear operator, so we can calculate it for (X_i) and then sum them up using alpha_i

$$E[X] = \sum_{i=1}^{n} \alpha_i E[X_i]$$

A.4(c) [4 points] Show that $Var[X] = \sum_{i=1}^{n} \alpha_i \left(\sigma_i^2 + \mu_i^2\right) - \mu^2$, where $\sigma_i^2 = Var[X_i]$, $\mu_i = E[X_i]$, and $\mu = E[X]$ (note that the μ^2 is outside the summation).

HINT: Use the fact that, for a random variable Y, $Var[Y] = E[Y^2] - [E(Y)]^2$ (it may be helpful at some point to re-arrange these terms).

$$\text{Var}[X] = E[X^2] - [E(X)]^2$$

$$\text{Var}[X] = E\left[\left(\sum_{i=1}^{n} \alpha_i X_i\right)^2\right] - \left[E\left(\sum_{i=1}^{n} \alpha_i X_i\right)\right]^2$$

$$\text{Var}[X] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j X_i X_j\right] - \left[E\left(\sum_{i=1}^{n} \alpha_i X_i\right)\right]^2$$

$$\text{Var}[X] = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j E[X_i X_j] - \left[E\left(\sum_{i=1}^{n} \alpha_i X_i\right)\right]^2$$

$$\text{Var}[X] = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j [E(X_i X_j) - E(X_i)E(X_j)] - \left[E\left(\sum_{i=1}^{n} \alpha_i X_i\right)\right]^2$$

$$\text{Var}[X] = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \text{Cov}[X_i, X_j] - \left[E\left(\sum_{i=1}^{n} \alpha_i X_i\right)\right]^2$$

Where _ij is the Kronecker delta, equal to 1 when i = j and 0 otherwise.

$$\text{Var}[X] = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j (\sigma_i^2 \delta_{ij} + \mu_i \mu_j) - \left[\sum_{i=1}^{n} \alpha_i E(X_i)\right]^2$$

$$\text{Var}[X] = \sum_{i=1}^{n} \alpha_i (\sigma_i^2 + \mu_i^2) - \left(\sum_{i=1}^{n} \alpha_i \mu_i\right)^2$$

$$\text{Var}[X] = \sum_{i=1}^{n} \alpha_i (\sigma_i^2 + \mu_i^2) - \mu^2$$

I got this answer from online but I think I understand how it works.

2 B Computational problems

2.1 B.1 Monte Carlo estimation

One really cool (and useful!) application of random variables is approximating integrals/the area under a curve. Consider $f(x) = sin(x)$ on the interval $0 \leq x \leq \pi$. Let's use uniform random variables to approximate the area under $f(x)$ on this interval. Note that this general idea is used often to solve really important but hard integrals.

B.1(a) [6 points] By hand, and using the `integrate()` function in R, calculate the true area under $f(x)$.

$$\int_0^\pi \sin(x)\, dx = 2$$

[ ]:

```
f <- function(x) sin(x)
answer <- integrate(f, lower = 0, upper = pi)
answer
```

```
2 with absolute error < 2.2e-14
```

B.1(b) [3 points] Generate $n = 5,000$ uniform random (x, y) coordinates in the rectangle $x \in [0, \pi]$, $y \in [0, 1]$.

[ ]:

```
n <- 5000

x <- runif(n, min = 0, max = pi)
y <- runif(n, min = 0, max = 1)

head(x)
head(y)
```

```
1. 0.40064343857879 2. 0.323111072770261 3. 0.619124877493399 4. 2.11129360437887
5. 0.0536899690655752 6. 1.1148128638591
```

```
1. 0.883794483030215 2. 0.985567841678858 3. 0.549567498732358 4. 0.264186913846061
5. 0.224520311923698 6. 0.698020716197789
```

B.1(c) [6 points] Calculate the proportion of points from B.1(b) that fall below $f(x)$ and use this proportion to approximate the area under $f(x)$.

[ ]:

```
f <- function(x) sin(x)
fx <- f(x)

prob_below_f <- sum(y <= fx) / n
prob_below_f

area <- prob_below_f  * pi
area
```

```
0.6454
```

2.02758389862685

The proportion below f(x) is around 63.5%. The estimated area is 1.994

B.1(d) [4 points] Find the absolute difference between our approximation and the true area calculated in B.1(a). How can we make this error smaller?

[ ]:

```r
abs(area-2)
```

0.0275838986268524

We can make this error smaller by using more data points.

2.2 B.2

For this problem, let's assume that the probability of an event E, $P(E)$, means something like "the degree to which you believe E".

B.2(a) [8 points] In the same window, plot the PDFs of four different Beta distributions:

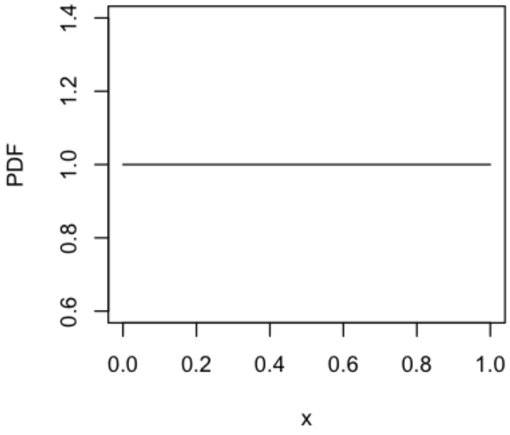
1. $\text{Beta}(1, 1)$
2. $\text{Beta}(0.5, 0.5)$
3. $\text{Beta}(5, 5)$
4. $\text{Beta}(1, 5)$

What is another name for 1**?**

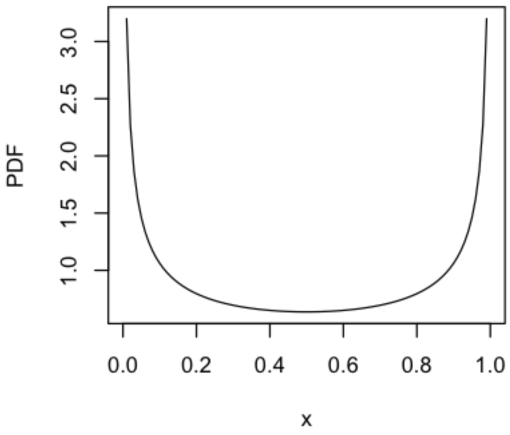
[ ]:

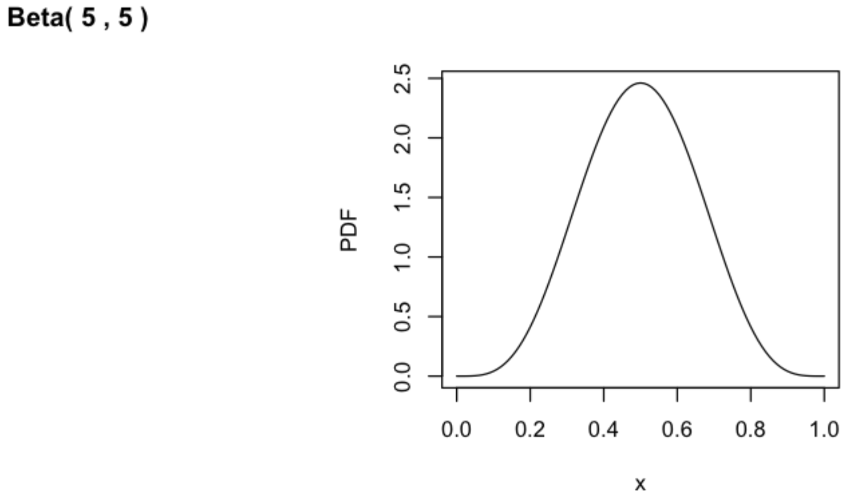
```r
x <- seq(0, 1, 0.01)

par(mfrow = c(2, 2))
params <- list(c(1, 1), c(0.5, 0.5), c(5, 5), c(1, 5))

for (i in 1:4) {
  plot.new()
  title(main = paste("Beta(", params[[i]][1], ",", params[[i]][2], ")"))
  curve(dbeta(x, shape1 = params[[i]][1], shape2 = params[[i]][2]),
        xlab = "x", ylab = "PDF")
}
```

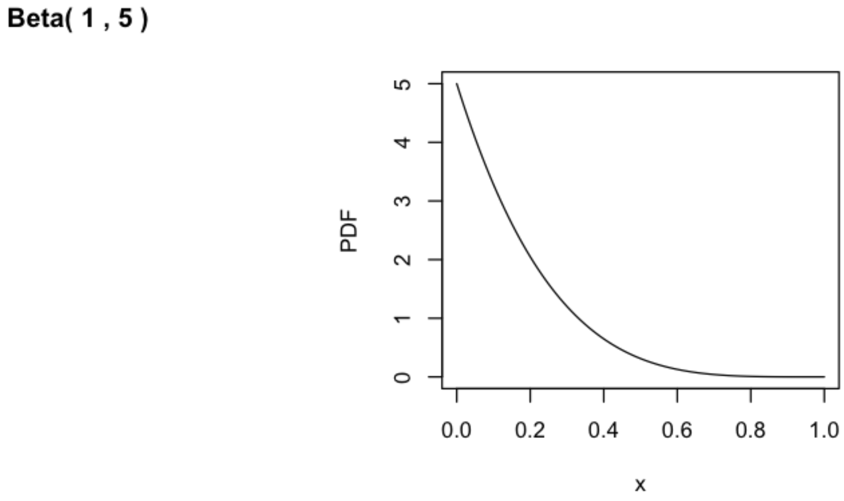
Beta(1 , 1)



Beta(0.5 , 0.5)







The value 1 is the shape parameter. It determines the shape or concentration of the distribution.

B.2(b) [8 points] Now, suppose your friend has a coin whose probabilities are unknown to you. Since the beta distribution has support $[0, 1]$, it is well-suited to model your beliefs about the probability of heads. Match the following descriptions with the beta distributions from the previous part.

i. I am pretty confident that this is a two-headed or two-tailed coin.

ii. I have no idea what the probability of heads is.

iii. I believe that the probability of heads is low.

iv. I believe that the coin is close to fair.

Assuming that 0 is tails and 1 is head:

i -> beta(0.5,0.5), the distribution has two modes at 0 and 1. This indicates a high likelihood that the coin is two-headed/tailed

ii -> beta(1,1), beta(1,1) indicates a lack of prior information or bias. Therefore you have no idea what the probability is.

iii -> beta(1,5), since most of the distribution is around 0, which is tails

iv -> beta(5,5), the distribution is centered at 0.5, which signifies a fair coin

2.3 B.3 Mixture distributions

Most of the distributions we've looked at so far have been unimodal, flat, or monotonically increasing/decreasing. Let's take a look at a class of distributions that are multimodal. You proved some results about these in problem **A.4**, but recall the definition:

Let $X_1, ..., X_n$ be continuous random variables with pdfs $f_{X_i}(x)$, $i = 1, ..., n$. Define a *mixture distribution* to be a random variable X with PDF

$$f_X(x) = \sum_{i=1}^{n} \alpha_i f_{X_i}(x)$$

where α_i are nonnegative real numbers such that $\sum_{i=1}^{n} \alpha_1 = 1$.

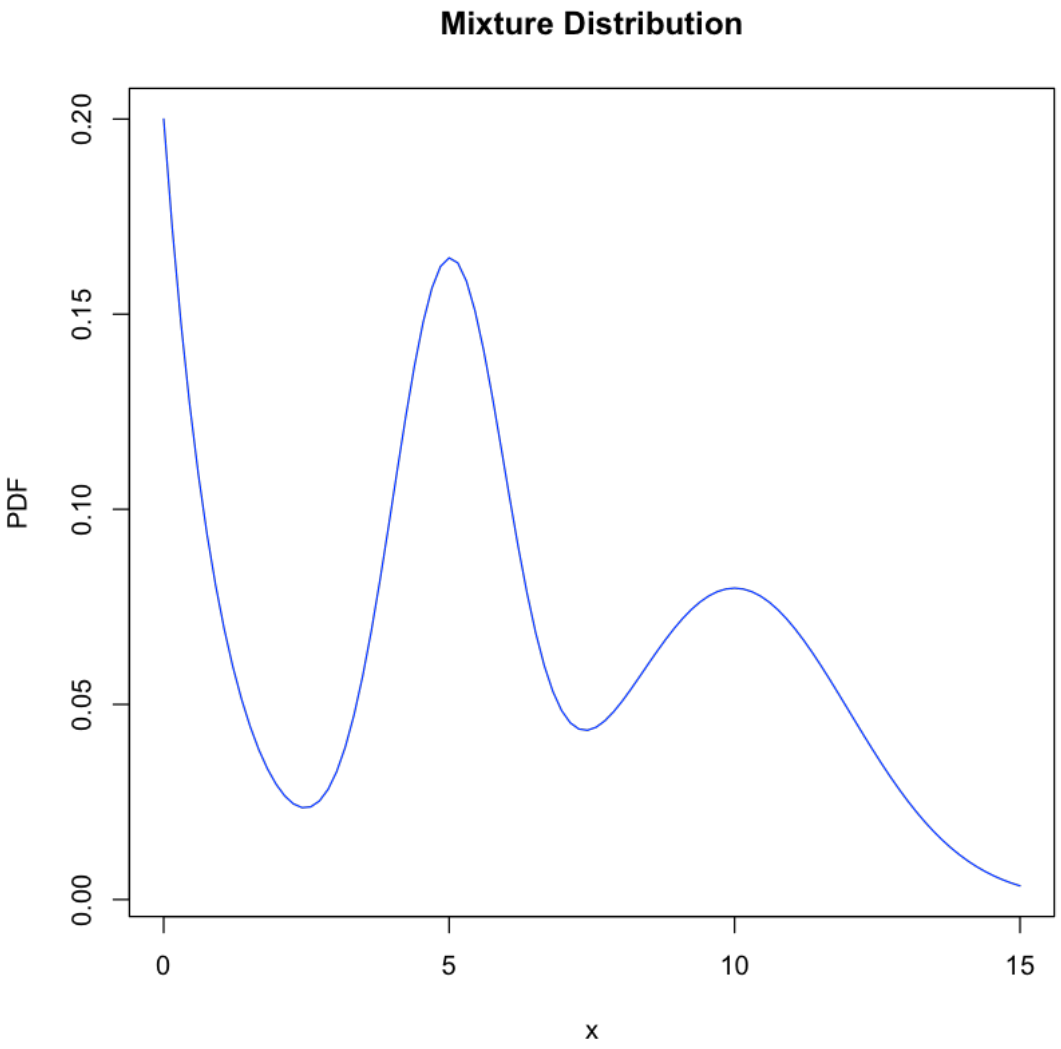
B.3(a) [8 points] For this question, let's create a specific mixture distribution and plot the pdf in R.

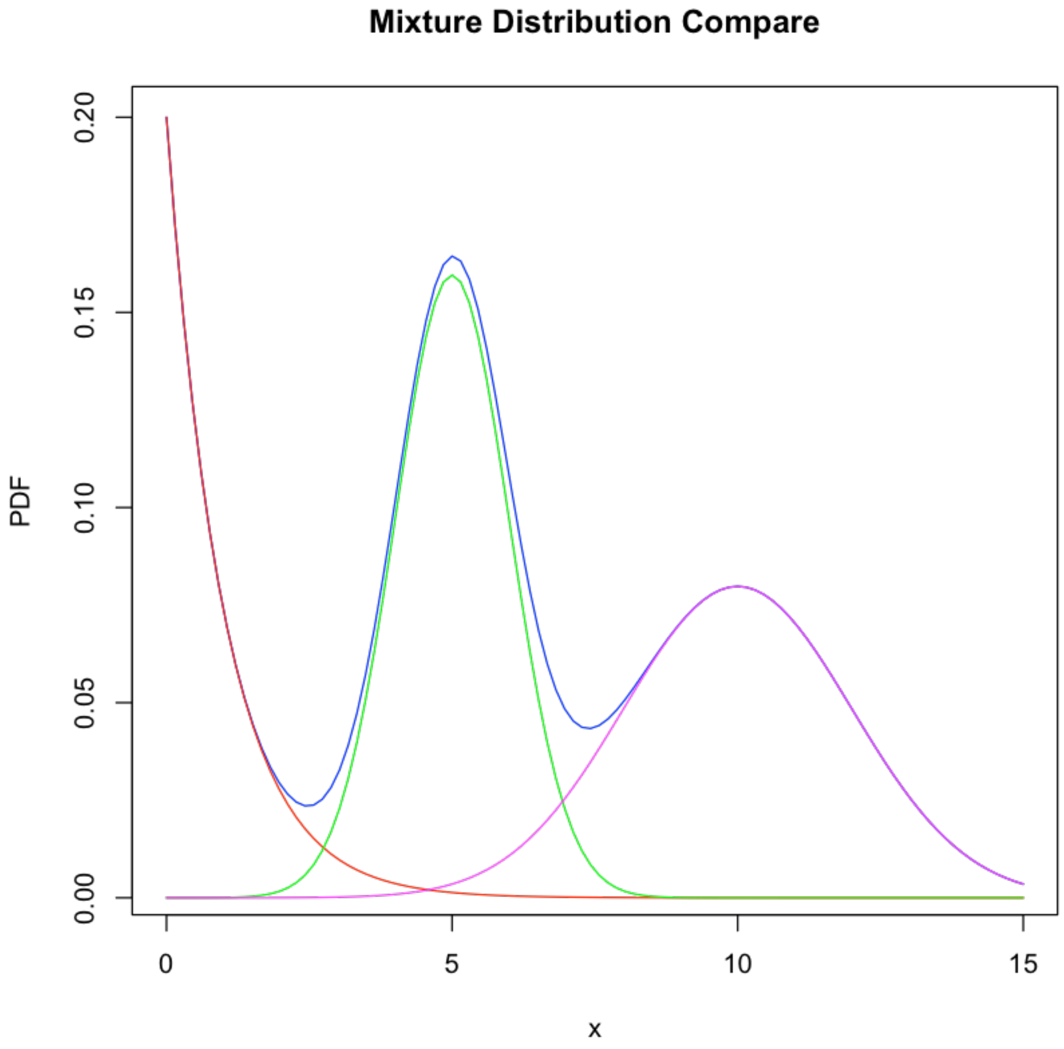
1. First, create a grid of `n = 100` `x` values evenly spaced between `0` and `15`.
2. Construct an `alpha` vector with entries `1/5`, `2/5`, and `2/5`.
3. At each value of `x`, calculate the following:
    - the PDF of an exponential with rate parameter `1`.
    - the PDF of a normal distribution with mean `5` and standard deviation `1`.
    - the PDF of a normal distribution with mean `10` and standard deviation `2`.
4. Construct the PDF of the mixture random variable using `alpha` and PDFs from 3.
5. Plot the pdf of the mixture distribution. What do you notice about it's shape? How does the shape relate to the shape of the original distributions?

[ ]:

```
x <- seq(0, 15, length = 100)

alpha <- c(1/5, 2/5, 2/5)

pdf_exp <- dexp(x, 1)
pdf_norm1 <- dnorm(x, 5, 1)
pdf_norm2 <- dnorm(x, 10, 2)

mix <- alpha[1] * pdf_exp + alpha[2] * pdf_norm1 + alpha[3] * pdf_norm2

plot(x, mix, type = 'l', col = 'blue', xlab = 'x', ylab = 'PDF',
     main = 'Mixture Distribution')
```

```r
plot(x, mix, type = 'l', col = 'blue', xlab = 'x', ylab = 'PDF',
     main = 'Mixture Distribution Compare')
lines(x, alpha[1] * pdf_exp, col = 'red')
lines(x, alpha[2] * pdf_norm1, col = 'green')
lines(x, alpha[3] * pdf_norm2, col = 'magenta')
```





The mixture has multiple modes/peaks centered at different x locations. When comparing the mixture with the three distributions individually, we can see that each peak in the mixture corresponds to the peak of the single distribution.

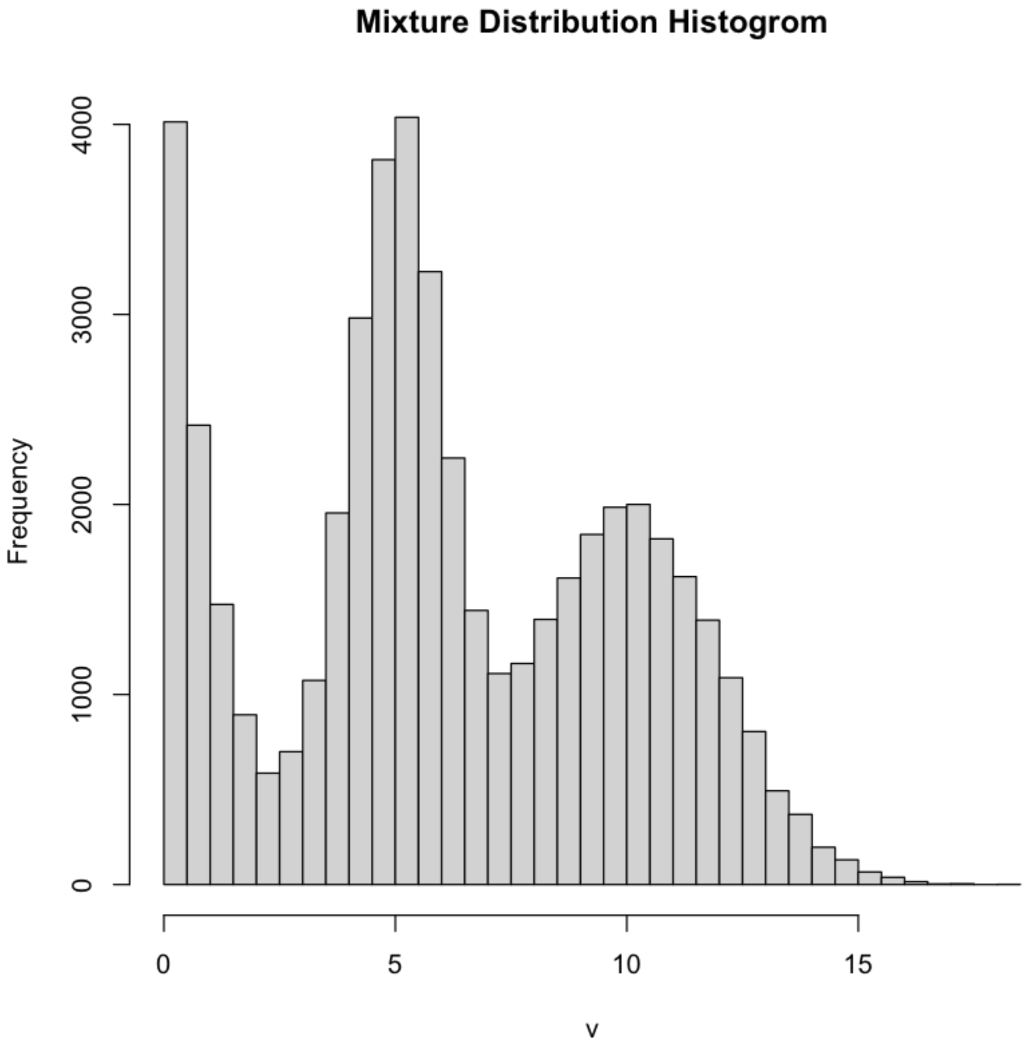
B.3(b) [8 points] In this part, generate $m = 50,000$ random numbers from the mixture distribution that we worked with in B.3(a), and store those numbers in a vector v. The algorithm below should help!

1. Generate a single random number from the continuous uniform distribution $U(0, 1)$. Call this number `u`.
2. Use `u` to select among the three distributions from B.3(a) (i.e., $Exp(1)$, $N(5, 1)$ or $N(10, 2)$). For example, generate a random number from $Exp(1)$ if `u < 1/5`.
3. Repeat this process `m = 50,000` times.

Then, create a histogram of `v`, and set `breaks = 30`. What do you notice about the distribution?

[ ]:

```
m <- 50000
v <- numeric(m)

# Generate random numbers from the mixture distribution
for (i in 1:m) {
  u <- runif(1)
  if (u < 1/5) {
    v[i] <- rexp(1, 1)
  } else if (u < 3/5) {
    v[i] <- rnorm(1, 5, 1)
  } else {
    v[i] <- rnorm(1, 10, 2)
  }
}
hist(v, breaks = 30, main = "Mixture Distribution Histogrom",
     xlab = "v", ylab = "Frequency")
```



This distribution looks similar to the one explored earlier.

B.3(c) [4 points] Compute the theoretical mean and variance for this mixture, i.e., $E(X)$ **and** $Var(X)$ **from A.4. Then, compare those values to the sample mean and variance of** `v`**. What do you notice?**

[ ]:

```
alpha <- c(1/5, 2/5, 2/5)
mu <- c(0, 5, 10)
sigma <- c(1, 1, 2)

E_X <- sum(alpha * mu)
Var_X <- sum(alpha * (sigma^2 + mu^2)) - E_X^2

print('Theoretical Mean:')
```

```
E_X
print('Theoretical Variance:')
Var_X

print('Real Mean:')
mean(v)
print('Real Variance:')
var(v)
```

```
[1] "Theoretical Mean:"
```

6

```
[1] "Theoretical Variance:"
```

16.2

```
[1] "Real Mean:"
```

6.19927678935329

```
[1] "Real Variance:"
```

14.0441951003205

The expected values are close to each other at around 6.

The variance values vary by 2.

I expect that if I increase the number of data points, the values will get closer.