Advanced

Models

Hw 1,

Part 2

Consider the following model the location of a tennis player's serve when serving to the right. Let

$$\mathcal{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_2\left(\begin{bmatrix} 29 \\ 16 \end{bmatrix}, \begin{bmatrix} 4 & 4 \\ 4 & 16 \end{bmatrix}\right)$$

where $N_2(v, \mathcal{C})$ is the bivariate normal distribution with mean vector v and covariance matrix $\mathcal{C}$ and X_1 and X_2 represent the lateral and depth locations of the landing point of the serve on the court. A serve is considered legal if the ball lands in the cross court service box, the area bounded by $18 \leq X_1 \leq 31.5$ and $0 \leq X_2 \leq 21$. The following R code can be used to generate an image of the tennis court.

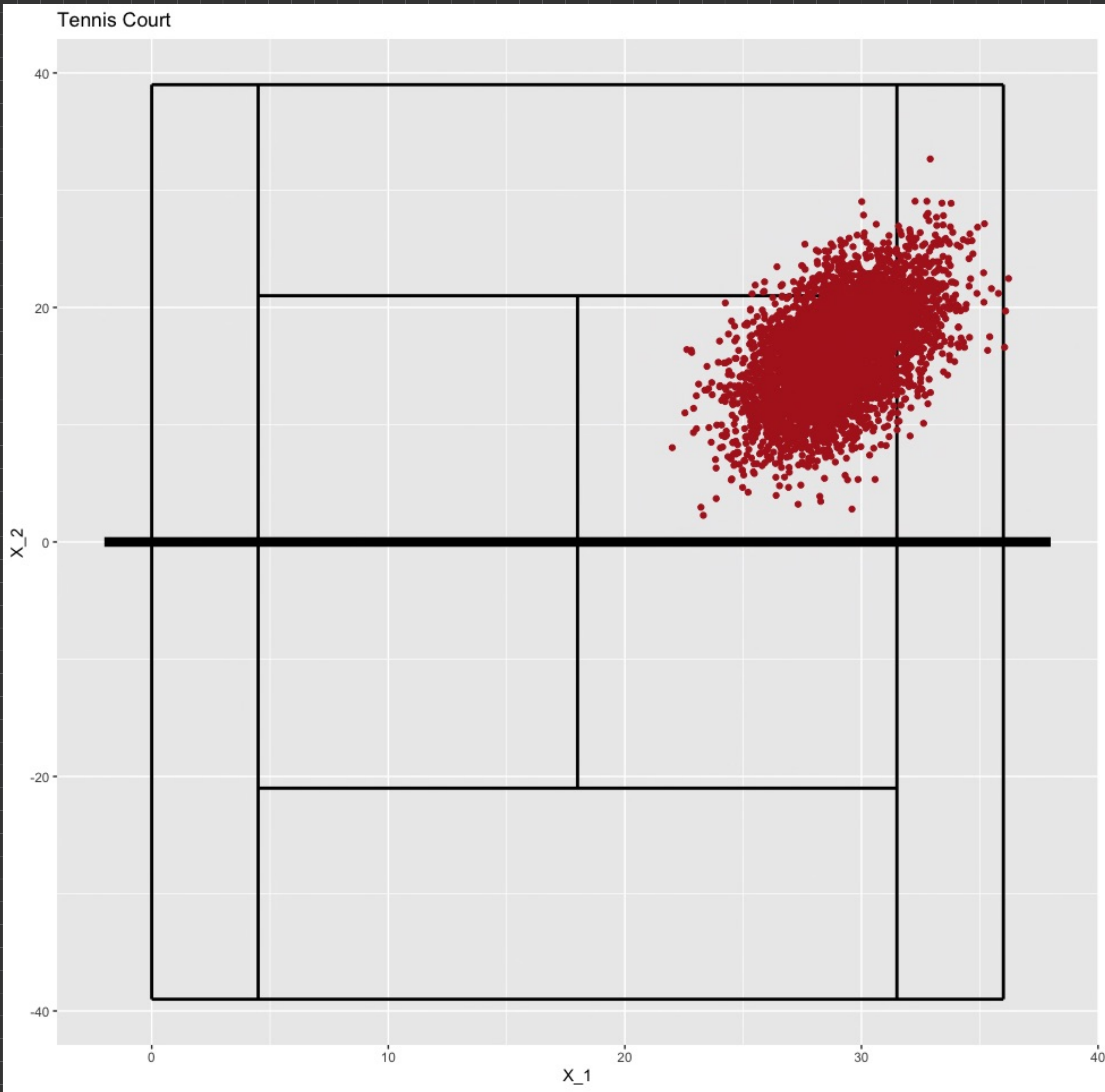
TennisCourtPlot.R

Question a

a. Generate 5,000 independent realizations of $\mathcal{X}$ and use ggplot to create a scatterplot of your simulated values of $\mathcal{X}$ over the provided tennis court plot.

```r
#Tennis Court Plot
#Importing ggplot2
library(ggplot2)

#Creates a data.frame object
tennisCourt = data.frame(x1 = c(0,4.5,18,31.5,36,0,4.5,4.5,0,-2),
                         x2 = c(0,4.5,18,31.5,36,36,31.5,31.5,36,38),
                         y1 = c(-39,-39,-21,-39,-39,39,21,-21,-39,0),
                         y2 = c(39,39,21,39,39,39,21,-21,-39,0),
                         width = c(rep(1,9),3))

#Creates a plot object called ggTennis
ggTennis = ggplot(tennisCourt) +
  geom_segment(aes(x = x1,y = y1,xend = x2,yend = y2),size = tennisCourt$width) +
  labs(x = "X_1",y = 'X_2',
       title = 'Tennis Court')

# ----------------------------------------------------------------------Question a-------------
# creating the covariance matrix  with the information given by the homework
m_sigma <-matrix(data=c(4,4,4,16),nrow=2,byrow=TRUE)
mean_v <- c(29,16)
# Creating 5000 points taken by the bivariate normal distribution
rand_points<- mvtnorm::rmvnorm(n=5000,mean=mean_v,sigma=m_sigma)
#Giving names to the columns of the data containing the 5000 points
colnames(rand_points)<-c("x","y")
#Making the data points as a dataframe
rand_points <- as.data.frame(x=rand_points)
#Data frame of the points to add
pointsToAdd = data.frame(x =rand_points$x, y = rand_points$y)
#Now we add in the points, and create a new object
#The geom_point function helps us create points. Note that we give it new data,
ggTennisWithPoints = ggTennis +
  geom_point(data = pointsToAdd,aes(x = x, y = y),color = 'firebrick')

#Let's see what we made
ggTennisWithPoints
```

Result:



Question 6

b. Using the model, what is the theoretical probability a serve from the player will be legal? Additionally, show how you can approximate this probability from the realizations of $\mathcal{X}$ and provide the numeric value of your approximation.

b) $P(\text{serve legal}) = P(18 \leq X_1 \leq 31.5, 0 \leq X_2 \leq 21)$

$$= \int_0^{21}\int_{18}^{31.5} \mathcal{N}\left(\begin{bmatrix}29\\16\end{bmatrix}, \begin{bmatrix}4 & 4\\4 & 16\end{bmatrix}\right) dx \approx \underline{0.82}$$

↳ to solve we use R

```r

# Using the multivariate normal distribution  function to calculate the theoretical probability.
#We use the mean vector and the covariance matrix given by the homework.
#We use the limits given by the exercise
th_prob<-mvtnorm::pmvnorm(lower=c(18,0),upper=c(31.5,21),mean=mean_v,sigma=m_sigma)

print("The theoretical probability is")
print(th_prob)

counter<- 0

#Count how many of the 5000 generated points are valid points
for (i in 1:5000)
{
  if  (rand_points$x[i]<=31.5  & rand_points$x[i]>=18   &   rand_points$y[i]>=0  & rand_points$y[i]<=21 )
  {
    counter <- counter + 1
  }
}

#The probability will be the number of events that are legal/ total events .
prob<- counter/5000

print("The  probability calculated by the simulation is")
print(prob)
```

Results

```
> print("The theoretical probability is")
[1] "The theoretical probability is"
> print(th_prob)
[1] 0.8236971
attr(,"error")
[1] 1e-15
attr(,"msg")
[1] "Normal Completion"
```

Theoretical Result

```
> print("The  probability calculated by the simulation is")
[1] "The  probability calculated by the simulation is"
> print(prob)
[1] 0.8342
```

Simulation Result

[redacted]

Question c

c. Say the player decides to evaluate their serves that land further to the right (positive X_1 direction). Given that the player examines their serves landing around $X_1 = 30.5$, what is the conditional distribution of X_2? What is the probability that these serves are legal (only considering depth, not width)? (Hint: Consider using the \textit{pnorm} function)

We want to find the conditional distribution $X_2 | X_1$. We can use the formulas learned at class.

We define

$$Y = \begin{bmatrix} X_2 \\ X_1 \end{bmatrix} \qquad \Sigma_Y = \left[\begin{array}{c|c} 16 & 4 \\ \hline 4 & 4 \end{array}\right]$$

$$\mu_1 = \mu_{Y_1} = \mu_{X_2} = 16$$

$$\mu_2 = \mu_{Y_2} = \mu_{X_1} = 29$$

$$\mu_{Y_1 | Y_2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2)$$

$$= 16 + 4 \cdot \frac{1}{4} (30.5 - 29)$$

$$= 16 + (1.5) = 17.5$$

$$\boxed{\mu_{Y_1|Y_2} = 17.5}$$

$$\Sigma_{Y_1|Y_2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

$$= 16 - 4 \left(\frac{1}{4}\right) 4 = 16 - 4 = 12$$

$$\boxed{\Sigma_{Y_1|Y_2} = 12}$$

Hence $X_2 | X_1 \sim N(17.5, 12)$

$$P(0 \le X_2 \le 21 \mid X_1 = 30.5)$$

$$= 0.843891$$

```r
#Calculated mean and standard deviation of the conditional distribution given x_1=30.5
mean_c <-17.5
sd_c <-12**0.5

# Probability of P(X_2 <= 21 | X_1 = 30.5)
p_21<-pnorm(21,mean = mean_c,sd =sd_c,lower.tail = TRUE, log.p = FALSE)
# Probability of P(X_2 <= 0 | X_1 = 30.5)
p_0<-pnorm(0, mean = mean_c,sd = sd_c,lower.tail = TRUE, log.p = FALSE)

# Probability of P(0<=X_2 <= 21 | X_1 = 30.5)
p_0_21<-p_21-p_0

print(p_0_21)
```

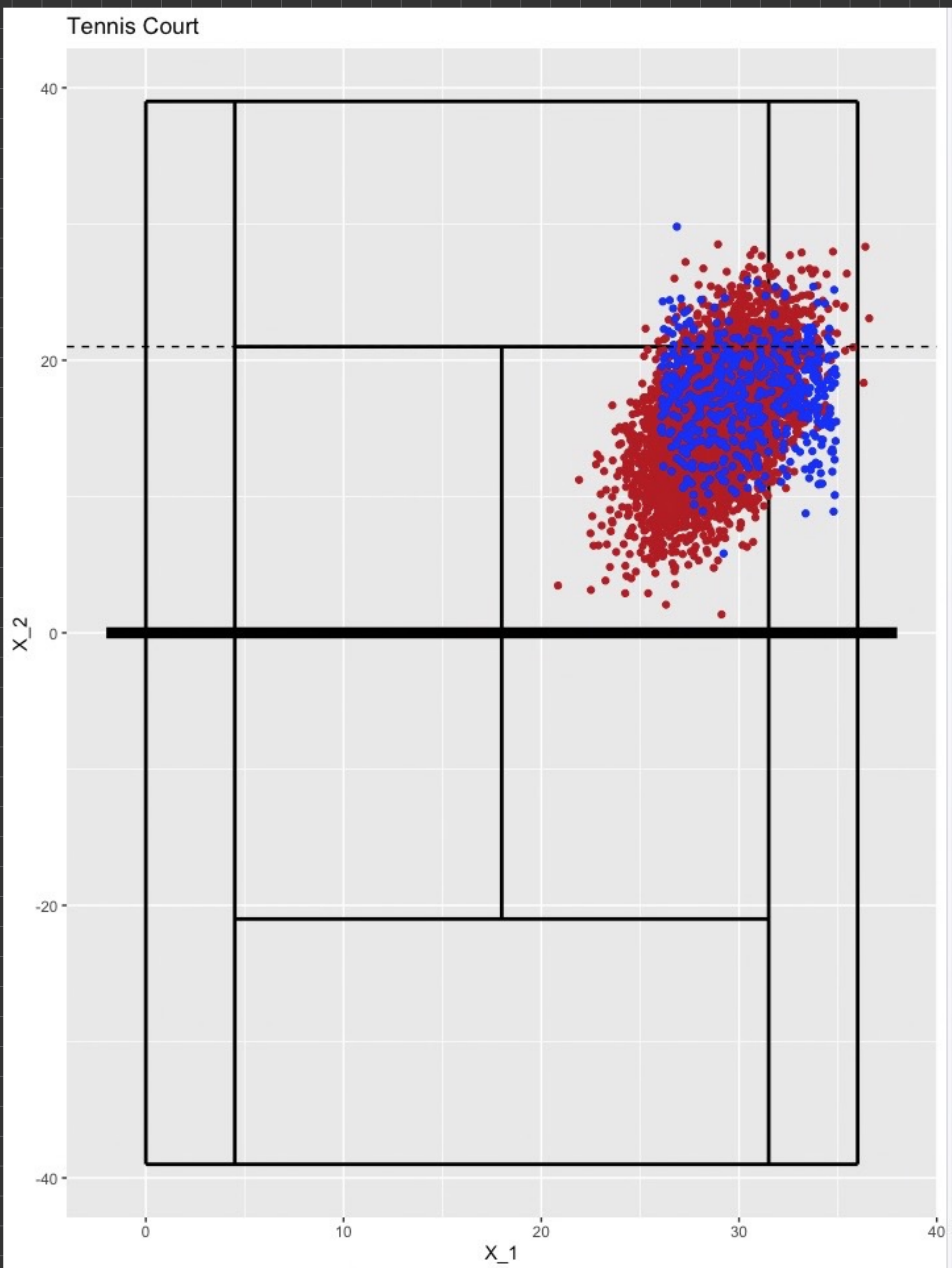
```
> print(p_0_21)
[1] 0.8438391
>
```

Result

Question d

d. Generate 500 realizations of X_2 from the conditional distribution found in part c. Create a new version of the scatterplot constructed for part a that includes plots of the 500 realizations of X_2 plotted as different color points with X_1 fixed at 30.5. Add a small amount of random noise to the X_1 component to reduce the effects of overplotting (consider using R's *jitter* function). Describe your results. How do your values of X_2 generated from the conditional distribution compare to the values generated directly from the original distribution?

```r
# ----------------------------------------------------------------------------Question d-----------------------------------------------------

#Calculated mean and standard deviation of the conditional distribution given x_1=30.5
mean_c <-17.5
sd_c <-12**0.5
# Create random X_2 coordinate points, using the mean and standard deviation calculated with the conditional distribution
rand_points_cond<- rnorm(n=500,mean=mean_c,sd=sd_c)

#Creating the random points, with the X_2 coordinate found with the conditional distribution, and giving the coordinate of
#X_1= 30.5

    # Using jitter function  to reduce the effects of overplotting
x_c<- jitter(rep(1, 500)*30.5,7.3)

rand_points_cond <- as.data.frame(x=rand_points_cond)
rand_points_cond2 <- cbind(x_c, rand_points_cond)

# Names to the data as x_c (x conditioned) and y_c (y conditioned)
colnames(rand_points_cond2)<-c("x_c","y_c")

# Conditioned points to be added
pointsToAdd_cond = data.frame(x =rand_points_cond2$x_c, y =rand_points_cond2$y_c)

#Adding, the random points of the original distribution + the new conditioned points in the tennis court
ggTennisWithPoints_wc = ggTennis +
  geom_point(data = pointsToAdd,aes(x = x, y = y),color = 'firebrick') +
  geom_point(data = pointsToAdd_cond,aes(x = x, y = y),color = 'blue') +geom_hline(yintercept=21,linetype=2)

#Let's see what we made
ggTennisWithPoints_wc
```



Result:

Once we know that $X_1 = 30.5$, the possible outcomes are now centered at higher depth (greater X_2)

The knowledge of $X_1 = 30.5$, have given us the information that is less likely to have events at lower depths (lower X_2 values)