ST-540 Assignment-3

Tilekbek Zhoroev

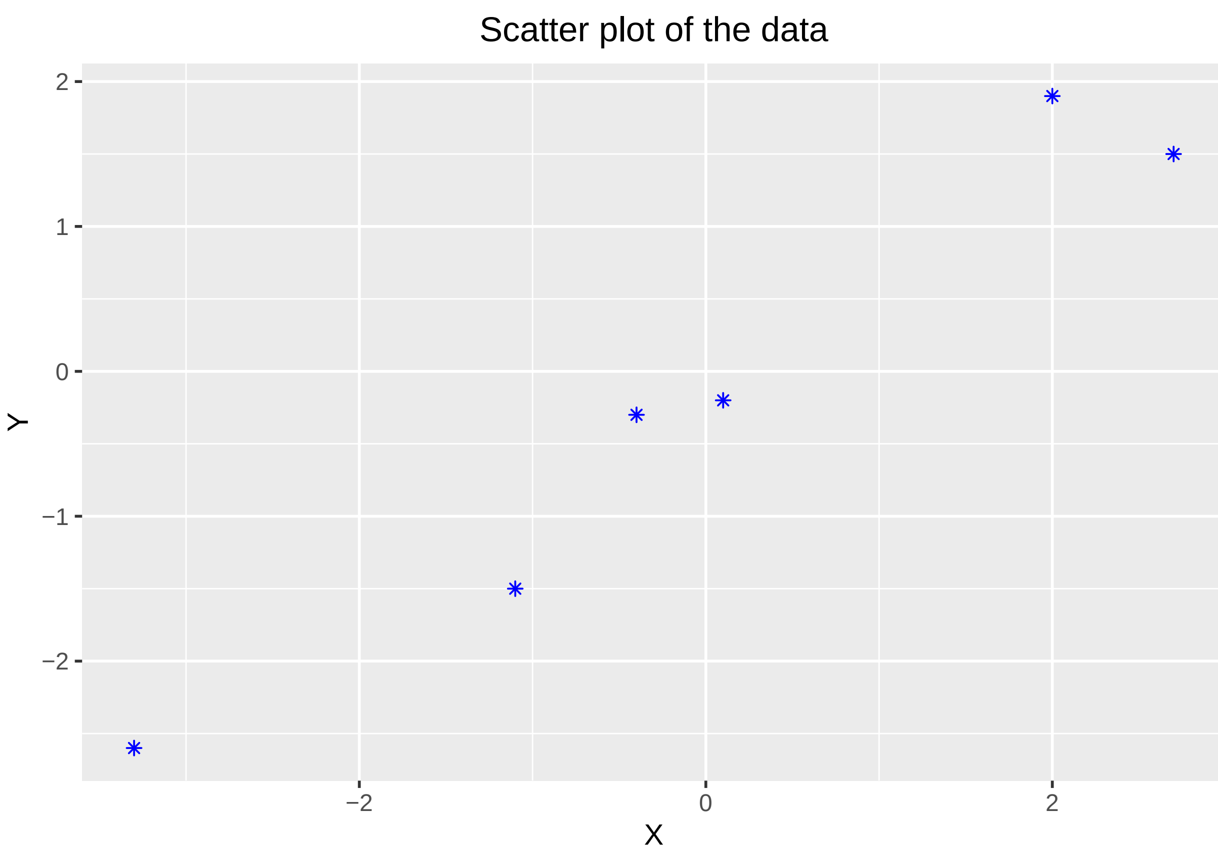
1/26/2022

Problem 1

Assume that $(X;Y)$ follow the bivariate normal distribution and that both X and Y have marginal mean zero and marginal variance one. We observe six independent and identically distributed data points: $(-3.3,-2.6),(0.1,-0.2),(-1.1,-1.5),(2.7,1.5),(2.0,1.9)$ and $(-0.4,-0.3)$. Make a scatter plot of the data and, assuming the correlation parameter ρ has a Uniform(-1; 1) prior, plot the posterior distribution of ρ.

Solution

```
df<-data.frame(X=c(-3.3, 0.1, -1.1, 2.7, 2.0, -0.4), Y=c(-2.6, -0.2, -1.5, 1.5, 1.9, -0.3))
df%>%
  ggplot(aes(x=X, y=Y))+
  geom_point(shape = 8, colour="blue")+
  theme(plot.title = element_text(hjust = 0.5))+
  ggtitle("Scatter plot of the data")
```



We have given that $\mu_x = \mu_y = 0$ and $\sigma_x = \sigma_y = 1$, the the joint distribution of X, Y with given ρ become

$$f(X,Y|\rho) = \frac{1}{2\pi\sqrt{1-p^2}} e^{-\frac{x^2+y^2-2xy\rho}{2(1-\rho^2)}}$$

Then using the fact that given data points are iid and Bayes' Theorem we obtained that

$$\begin{aligned}
P(\rho|(X_1,Y_1),...,(X_2,Y_2)) &\propto \prod_{i=1}^{6} P(X_i,Y_i|\rho)P(\rho) &(1)\\
&= \prod_{i=1}^{6} \frac{1}{2\pi\sqrt{1-p^2}} exp(-\frac{x_i^2+y_i^2-2x_iy_i\rho}{2(1-\rho^2)})U(-1,1) &(2)\\
&\propto \frac{1}{(\sqrt{1-p^2})^3} exp(-\frac{38.56-2*18.18\rho}{2(1-\rho^2)}) &(3)
\end{aligned}$$

where we used

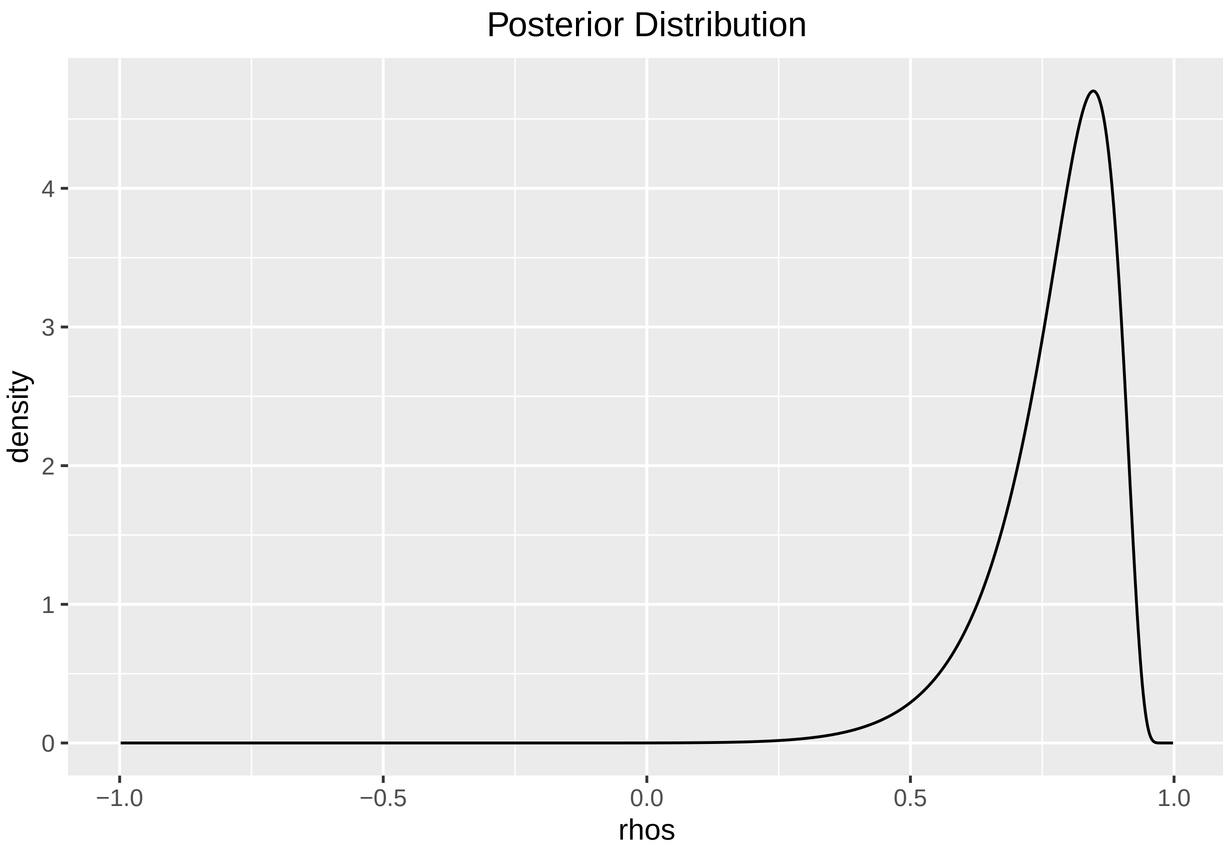
```
sum(df^2)
```

```
## [1] 38.56
```

```
sum(df$X*df$Y)
```

```
## [1] 18.18
```

```
post<-function(rho){1/(1-rho^2)^3*exp(-(38.56-rho*36.36)/(2-2*rho^2))}
# normalization factor
int_val<-integrate(post, lower = -1, upper = 1)$value
# since rho is between -1, 1 take sequence between this points,
rhos <- seq(-1,1, length.out=1000)
# posterior distribution values
data.frame(rhos, density = sapply(rhos, post)/int_val)%>%
  filter(!is.na(density))%>%
  ggplot(aes(x=rhos,y=density))+
  geom_line()+theme(plot.title = element_text(hjust = 0.5))+
  ggtitle("Posterior Distribution")
```



Problem 2

The normalized difference vegetation index (NDVI) is commonly used to classify land cover using remote sensing data. Hypothetically, say that NDVI follows a Beta(25; 10) distribution for pixels in a rain forest, and a Beta(10; 15) distribution for pixels in a deforested area now used for agriculture. Assuming about 10% of the rain forest has been deforested, your objective is to build a rule to classify individual pixels as deforested based on their NDVI.

(a) Plot the PDF of NDVI for forested and deforested pixels, and the marginal distribution of NDVI averaging over categories.

(b) Give an expression for the probability that a pixel is deforested given its NDVI value, and plot this probability by NDVI.

(c) You will classify a pixel as deforested if you are at least 90% sure it is deforested. Following this rule, give the range of NDVI that will lead to a pixel being classiffed as deforested.

Solution

(a) Let X = Rain forest area, Y = deforested area. Let θ = NDVI averaging. We have given that

$$\theta|X \sim Beta(\alpha = 25, \beta = 10)$$

$$\theta|Y \sim Beta(\alpha = 10, \beta = 15)$$

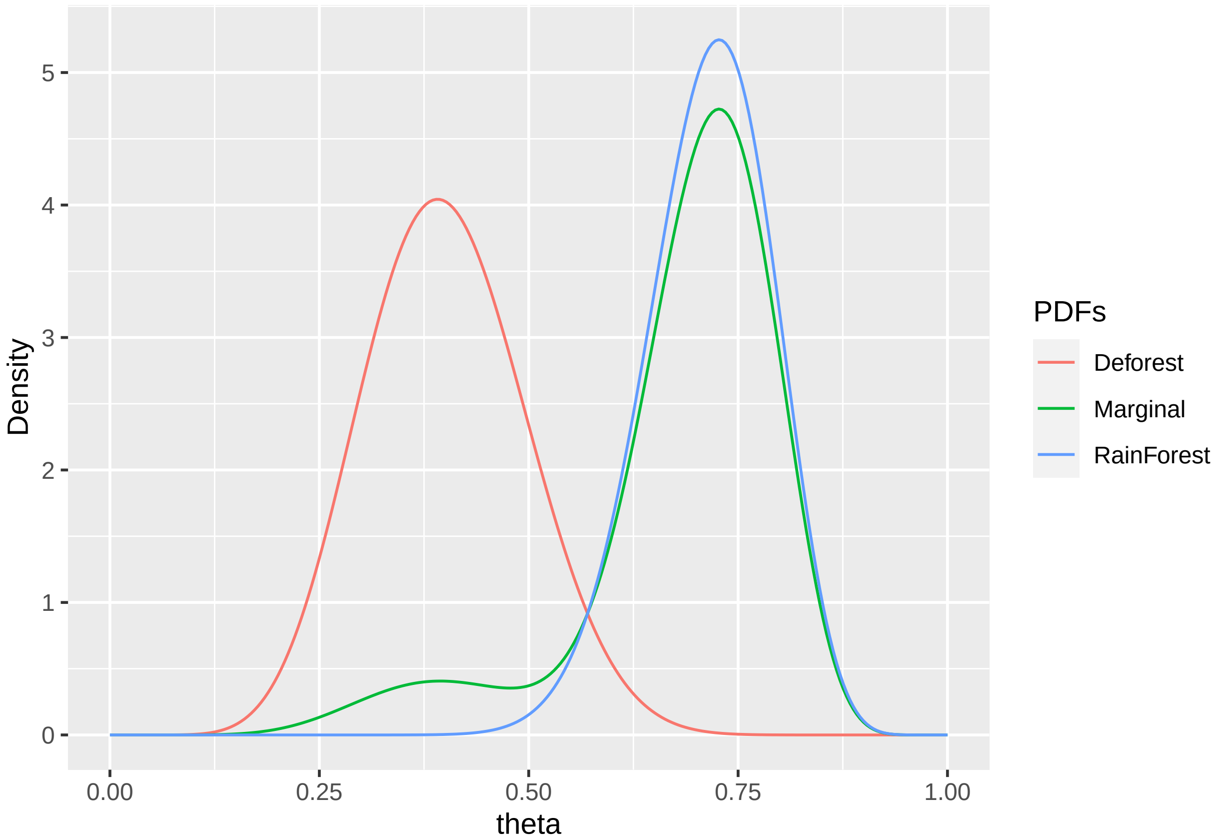
and

$$P(X) = 0.9, \qquad P(Y) = 0.1$$

Then using conditional probability the marginal distribution of NDVI averaging

$$\begin{aligned} f(\theta) &= f(\theta|X)P(X) + f(\theta|Y)P(Y) && (4) \\ &= .9Beta(25, 10) + .1Beta(10, 15) && (5) \end{aligned}$$

```
# since beta distribution is between 0 and 1 we take value at [0,1]
theta = seq(from = 0, to = 1, length.out = 250)
data.frame(theta, RainForest = sapply(theta, dbeta, 25,10),
           Deforest = sapply(theta, dbeta, 10,15))%>%
  mutate(Marginal = .9*RainForest +.1*Deforest)%>%
  gather(PDFs,Density, RainForest, Deforest,Marginal) %>%
  ggplot(aes(x = theta, y = Density, colour = PDFs ))+
  geom_line()
```

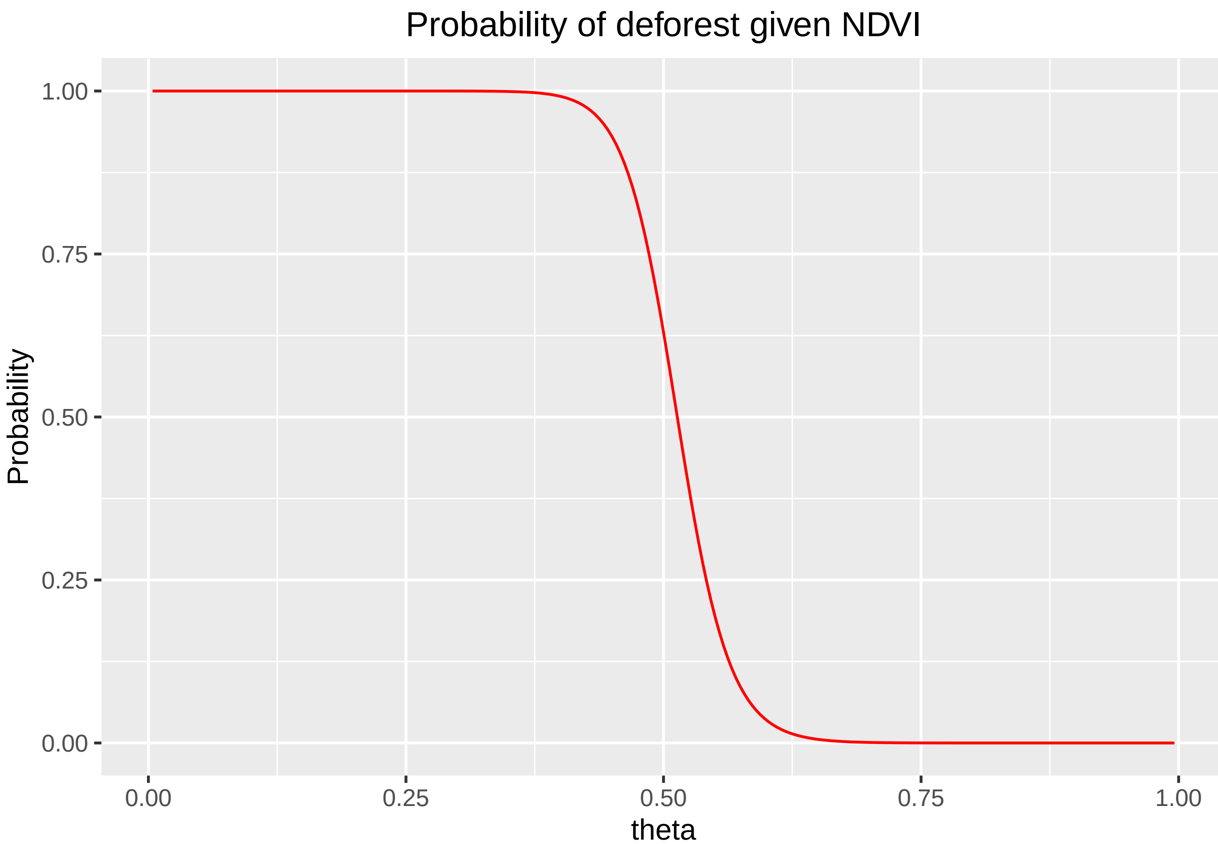


(b) Using Bayes' Rule we have

$$P(Y|\theta) = \frac{P(\theta|X)P(Y)}{P(\theta)} \tag{6}$$

$$= \frac{.1Beta(10,15)}{.9Beta(25,10) + .1Beta(10,15)} \tag{7}$$

```
data.frame(theta, num = .1*sapply(theta, dbeta,10,15),
           denom = .9*sapply(theta, dbeta,25,10) +.1*sapply(theta, dbeta,10,15))%>%
  mutate(Probability = num/denom)%>%
  filter(!is.na(Probability))%>%
  ggplot(aes(x=theta, y=Probability))+
  geom_line(colour = "red")+theme(plot.title = element_text(hjust = 0.5))+
  ggtitle("Probability of deforest given NDVI")
```



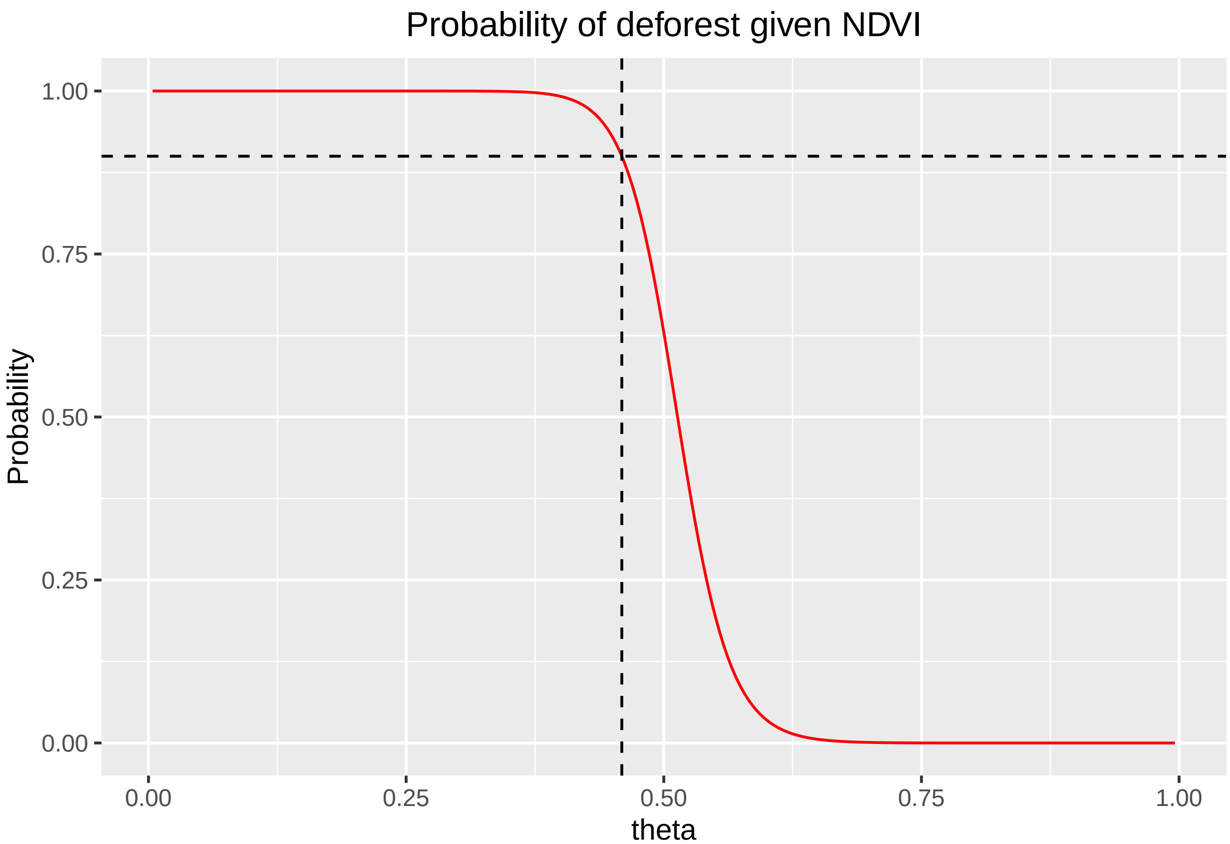
(c)

```
# since probability at end points are not defined I added machine epsilon to the left
# end point and subtracted it from upper end point of the interval [0,1].
uniroot(function(x) (.1*dbeta(x,10,15)/(.1*dbeta(x,10,15)+.9*dbeta(x,25,10))-.9),
        lower = 0+1e-16, upper = 1-1e-16, tol = 1e-9)$root
```

```
## [1] 0.4592829
```

Next, let us verify it graphically,

```
data.frame(theta, num = .1*sapply(theta, dbeta,10,15),
           denom = .9*sapply(theta, dbeta,25,10) +.1*sapply(theta, dbeta,10,15))%>%
  mutate(Probability = num/denom)%>%
  filter(!is.na(Probability))%>%
  ggplot(aes(x=theta, y=Probability))+
  geom_line(colour = "red")+theme(plot.title = element_text(hjust = 0.5))+
  ggtitle("Probability of deforest given NDVI")+
  geom_hline(yintercept=.9, linetype="dashed", color = "black")+
  geom_vline(xintercept=0.4592829, linetype="dashed", color = "black")
```



Problem 3

The table below has the overall free throw proportion and results of free throws taken in pressure situations, defined as "clutch" https://stats.nba.com/, for ten National Basketball Association players (those that received the most votes for the Most Valuable Player Award) for the 2016-2017 season. Since the overall proportion is computed using a large sample size, assume it is fixed and analyze the clutch data for each player separately using Bayesian methods. Assume a uniform prior throughout this problem

Player	Overall proportion	Clutch makes	Clutch attempts
Russell Westbrook	0.845	64	75
James Harden	0.847	72	95
Kawhi Leonard	0.880	55	63
LeBron James	0.674	27	39
Isaiah Thomas	0.909	75	83
Stephen Curry	0.898	24	26
Giannis Antetokounmpo	0.770	28	41
John Wall	0.801	66	82
Anthony Davis	0.802	40	54
Kevin Durant	0.875	13	16

(a). Describe your model for studying the clutch success probability including the likelihood and prior.

(b). Plot the posteriors of the clutch success probabilities.

(c). Summarize the posteriors in a table.

(d). Do you find evidence that any of the players have a different clutch percentage than overall percentage?

(e). Are the results sensitive to your prior? That is, do small changes

Solution

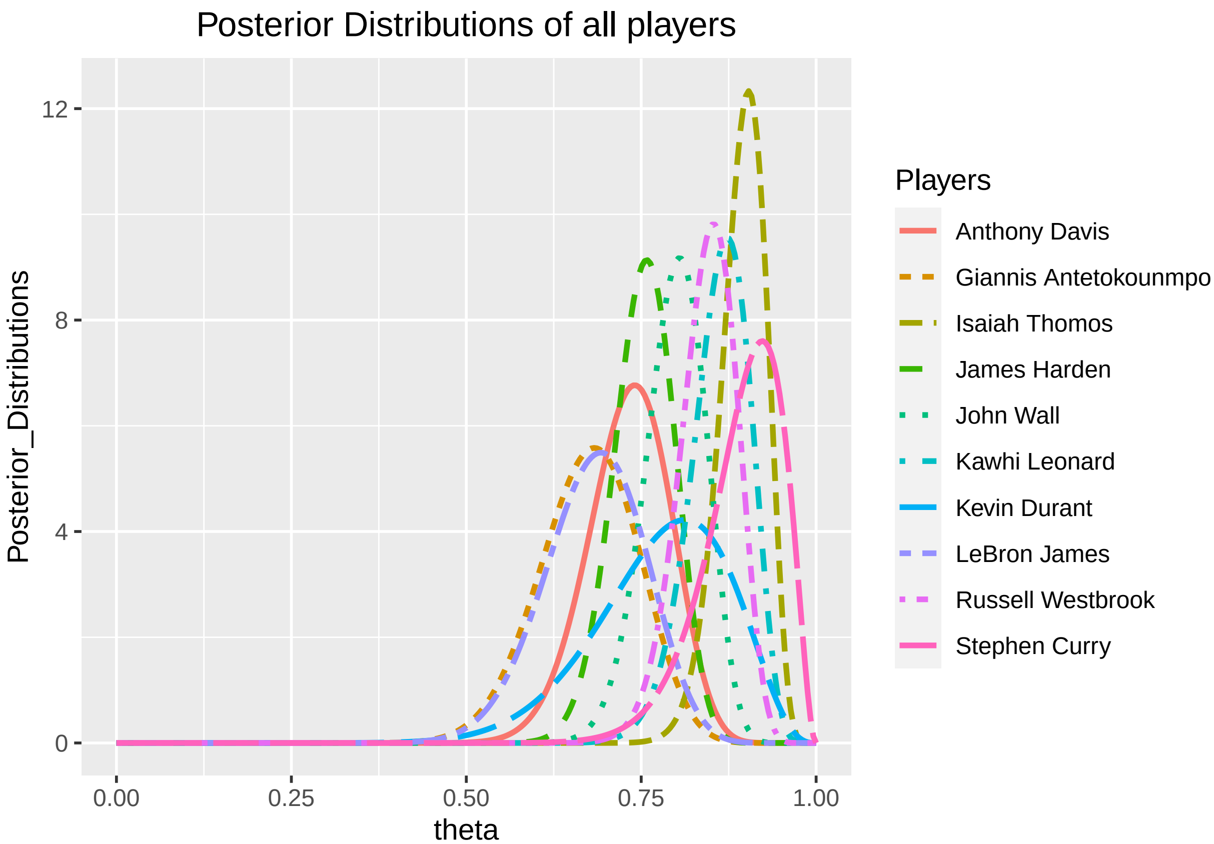
(a). Let θ be the probability of the success cluth and we already given that the priors are uniform, or $Beta(1, 1)$. Since we have given success/failure data we assume the likelihood distributions as Binomial distribution. From lecture notes we know that Binomial distribution is conjugate of Beta distribution, hence the posterior distributions are become Beta distribution,

$$\theta|Y_i \propto Beta(Y_i + 1, N_i - Y_i + 1)$$

where N_i and Y_i are clutch attempts and clutch makes numbers of each players respectively.

(b)

```
Ys = c(64,72,55,27,75,24,28,66,40,13)
Ns = c(75,95,63,39,83,26,41,82,54,16)
As = Ys+1
Bs = Ns - Ys + 1
names = c("Russell Westbrook","James Harden","Kawhi Leonard","LeBron James",
"Isaiah Thomos","Stephen Curry","Giannis Antetokounmpo","John Wall",
"Anthony Davis","Kevin Durant")
theta = seq(from = 0, to = 1, length.out = 250)
dt<-data.frame(theta)
for(i in 1:10){
   dt[names[i]]= sapply(theta, dbeta, As[i], Bs[i] )
}
dt%>%
pivot_longer(cols = "Russell Westbrook":"Kevin Durant", names_to = "Players",
             values_to = "Posterior_Distributions")%>%
  ggplot(aes(x=theta, y = Posterior_Distributions, color = Players, linetype = Players))+
  geom_line(size=1)+theme(plot.title = element_text(hjust = 0.5))+
  ggtitle("Posterior Distributions of all players")
```



(c) To summarize posterior distribution we give mean, standard deviation and 95% credible interval. We can use direct mean and standard deviation of beta distribution. Here we applied Monte Carlo sampling.

```
table<-data.frame(name=names, alpha=As,beta=Bs)
for(i in 1:10){
 theta<-rbeta(10^5, As[i], Bs[i])
 table[i,4]<-mean(theta)
 table[i,5]<-sd(theta)
 table[i,6]<-qbeta(0.025, As[i], Bs[i])
 table[i,7]<-qbeta(0.975, As[i], Bs[i])
}
colnames(table)<-c("Name", "Alpha", "Beta", "Means", "Standard Deviation", "2.5 %", "97.5 %")
knitr::kable(table)
```

Name	Alpha	Beta	Means	Standard Deviation	2.5 %	97.5 %
Russell Westbrook	65	12	0.8443824	0.0410775	0.7557660	0.9156623
James Harden	73	24	0.7523346	0.0435500	0.6625065	0.8327883
Kawhi Leonard	56	9	0.8612663	0.0425903	0.7684737	0.9336259
LeBron James	28	13	0.6827449	0.0716638	0.5346837	0.8142710
Isaiah Thomos	76	9	0.8941768	0.0332732	0.8209403	0.9498226
Stephen Curry	25	3	0.8928859	0.0576465	0.7571017	0.9764725
Giannis Antetokounmpo	29	14	0.6746615	0.0707943	0.5291394	0.8043320
John Wall	67	17	0.7971269	0.0437176	0.7059140	0.8759195
Anthony Davis	41	15	0.7322044	0.0585706	0.6099716	0.8386204
Kevin Durant	14	4	0.7773946	0.0953333	0.5656821	0.9318923

(d) To identify the differnce we obtain the posterior p-values,

$$p_{values} = P(\theta > \hat{\theta}|Y_i)$$

where $\theta \propto Beta(Y_i + 1, N_i - Y_i + 1)$.

```
theta_hat = c(0.845,0.847,0.880,0.674,0.909,0.898, 0.770,0.801,0.802,0.875)
p_vals<-data.frame(names)
for(i in 1:10){
  p_vals[i,2]<-pbeta(theta_hat[i], As[i], Bs[i], lower.tail = FALSE)
}
colnames(p_vals)<-c("Names", "Posterior p-values")
knitr::kable(p_vals)
```

Names	Posterior p-values
Russell Westbrook	0.5207121
James Harden	0.0090208
Kawhi Leonard	0.3597559
LeBron James	0.5645547
Isaiah Thomos	0.3553753
Stephen Curry	0.5293229
Giannis Antetokounmpo	0.0833433
John Wall	0.4907634
Anthony Davis	0.1133361
Kevin Durant	0.1542541

According to the table, it shows that J.Hardon and G.Antetokounmpo have relative small p-values(<0.1), which indicates huge discrepancy between the overall percentage and the posterior derived from data.

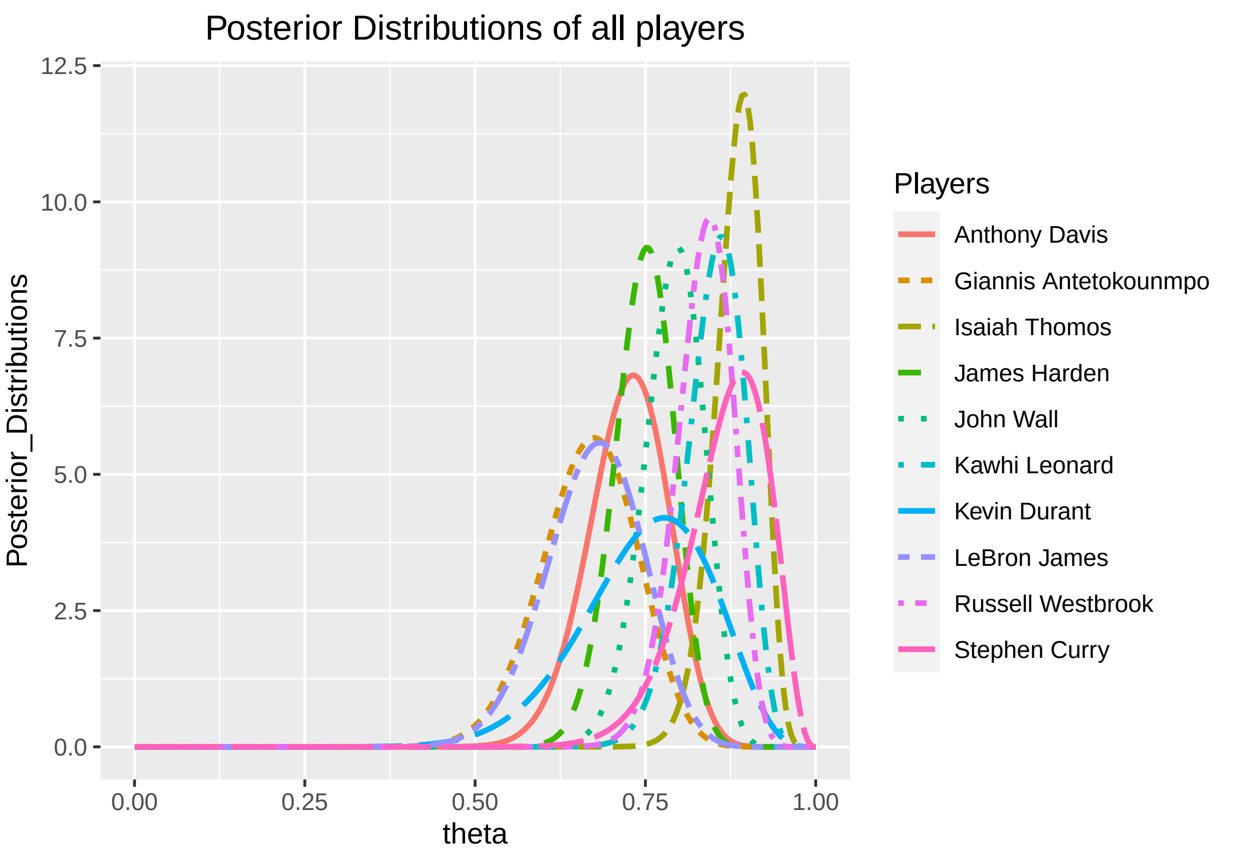
(e) Let us change prior slightly i.e. from $Beta(1, 1)$ to $Beta(2, 2)$ to check the sensitivity.

```
As_new <- Ys+2
Bs_new <-Ns-Ys+2
table<-data.frame(name=names, alpha=As_new,beta=Bs_new)
for(i in 1:10){
 theta_new<-rbeta(10^5, As_new[i], Bs_new[i])
 table[i,4]<-mean(theta_new)
 table[i,5]<-sd(theta_new)
 table[i,6]<-qbeta(0.025, As_new[i], Bs_new[i])
 table[i,7]<-qbeta(0.975, As_new[i], Bs_new[i])
}
colnames(table)<-c("Name", "Alpha", "Beta", "Means", "Standart Deviation", "2.5 %", "97.5 %")
knitr::kable(table)
```

Name	Alpha	Beta	Means	Standart Deviation	2.5 %	97.5 %
Russell Westbrook	66	13	0.8354028	0.0415515	0.7466787	0.9081616
James Harden	74	25	0.7476611	0.0434094	0.6578697	0.8276137
Kawhi Leonard	57	10	0.8506581	0.0432283	0.7568587	0.9248754
LeBron James	29	14	0.6742529	0.0707455	0.5291394	0.8043320
Isaiah Thomos	77	10	0.8851093	0.0338615	0.8106145	0.9428089
Stephen Curry	26	4	0.8664027	0.0609580	0.7264848	0.9611052
Giannis Antetokounmpo	30	15	0.6668783	0.0693853	0.5242186	0.7950827
John Wall	68	18	0.7909142	0.0436536	0.6992069	0.8693979
Anthony Davis	42	16	0.7242793	0.0581066	0.6033725	0.8302708

Name	Alpha	Beta	Means	Standart Deviation	2.5 %	97.5 %
Kevin Durant	15	5	0.7498563	0.0943931	0.5443469	0.9085342

```
theta = seq(from = 0, to = 1, length.out = 250)
dt<-data.frame(theta)
for(i in 1:10){
   dt[names[i]]= sapply(theta, dbeta, As_new[i], Bs_new[i] )
}
dt%>%
pivot_longer(cols = "Russell Westbrook":"Kevin Durant", names_to = "Players",
             values_to = "Posterior_Distributions")%>%
  ggplot(aes(x=theta, y = Posterior_Distributions, color = Players, linetype = Players))+
  geom_line(size=1)+theme(plot.title = element_text(hjust = 0.5))+
  ggtitle("Posterior Distributions of all players")
```



From table and posterior distribution plot we observe that, posterior statistics are not changed much. Hence, we conclude that posterior distribution is not sensitive to choice of priors.

Problem 4

The Major League Baseball player Reggie Jackson is known as "Mr. October" for his outstanding performances in the World Series (which takes place in October). Over his long career he played in 2820 regular-season games and hit 563 home runs in these games (a player can hit 0, 1, 2, ... home runs in a game). He also played in 27 World Series games and hit 10 home runs in these games. Assuming uninformative conjugate priors, summarize the posterior distribution of his home-run rate in the regular season and World Series. Is there sufficient evidence to claim that he performs better in the World Series?

Solution

Let λ_1 be the proportion of hits to total regular-season games and λ_2 be the proportion of hits to total World Series games. Then assume that prior is $\lambda_1 \propto \text{Gamma}(.1,.1)$ and $\lambda_2 \propto \text{Gamma}(.1,.1)$. Moreover, the likelihood function is Poisson distribution, then the posterior distribution become Gamma distribution

```
S <- 100000
a <- b <- .1 # uninformative prior
N1 <- 2820
N2 <- 27
Y1 <- 563
Y2 <- 10
# MC samples
lambda1 <- rgamma(S,Y1+a,N1+b)
lambda2 <- rgamma(S,Y2+a,N2+b)
# Prob(|data)
mean(lambda2>lambda1)
```

```
## [1] 0.95244
```

Probability of he performs better on World Series games than regular - season games is 0.9537 and it's sufficiently large.