Assignment 5

AUTHOR
Joaquin Peraza & Nicolas Giordano

Data model:

$$[y_{ij}|z_i] = \begin{cases} 0 \text{ if } z_i = 0, \\ Bern(a_j) \text{ if } z_i = 1, \end{cases}$$

Process model:

$$[Zi|p] = Bern(p)$$

Prior:

$[p] = Beta(\alpha, \beta)$

$\alpha = 1 \;\; \beta = 1$

Rapid test true positive rate: $[a_1] = 0.80$

PCR true positive rate: $[a_2] = 0.99$

*Zi -> The actual bat that was tested

$$[p, zi|y_{11}, y_{12}] = \frac{[y_{11}|p, z_i] * [y_{12}|p, z_i] * [z|p][p]}{\sum_{zi=0}^{1} \int_0^1 [y_{11}|p, z_i] * [y_{12}|p, z_i] * [z|p][p] \; dp}$$

$$[p|y_{11}, y_{12}] = \sum_{zi=0}^{1} [p, zi|y_{11}, y_{12}]$$

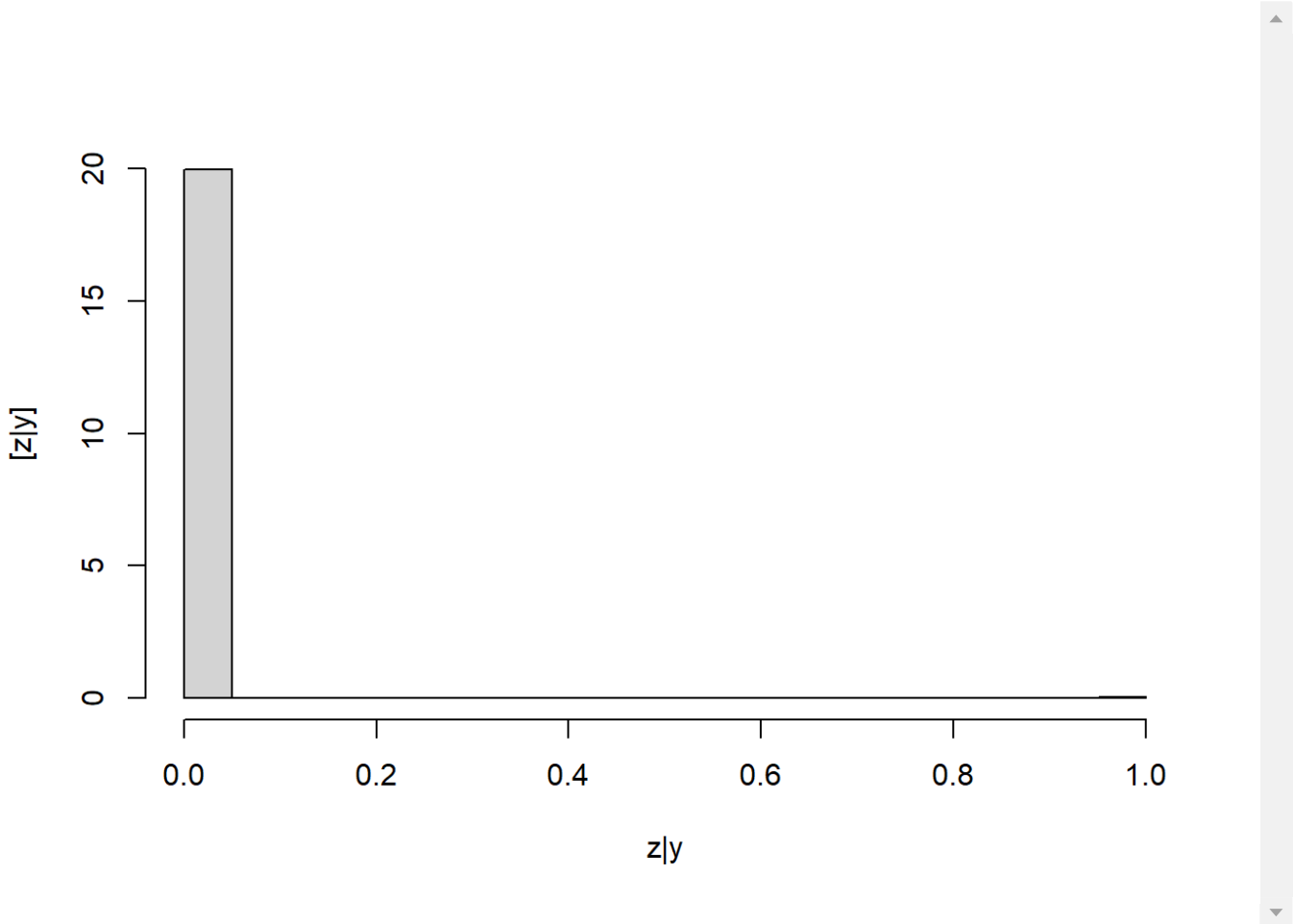
$$[z_i|y_{11}, y_{12}] = \int_0^1 [p, z_i|y_{11}, y_{12}] \; dp$$

```
set.seed(401)

# Data
y1 <- 0 # Results from rapid test
y2 <- 0 # Results from PCR test


# Hyperparameters for priors
alpha.p <- 1
beta.p <- 1

# Fixed (assumed known) values of true positive rate
a1 <- 0.80
a2 <- 0.99


# Preliminary MCMC stuff
K <- 50000
```

```
samples <- matrix(,K,2)
colnames(samples) <- c("z1","p")
p.initial <- 0.01


# Gibbs sampler
for(k in 1:K){

  # Take one draw from the full conditional of z1 (i.e., [z1|p,y1,y2])
  p <- ifelse(k==1,p.initial,p)
  p.tilde <- (p*(1-a1)*(1-a2))/((p*(1-a1)*(1-a2))+1-p)
  z1 <- ifelse(min(c(y1,y2))==0,rbinom(1,1,p.tilde),1)

  # Take one draw from the full conditional of p (i.e., [p|z1,y1,y2])
  p <- rbeta(1,alpha.p+z1,beta.p+1-z1)

  # Save samples
  samples[k,] <- c(z1,p)
}

# Discard burn-in interval
samples <- samples[10001:K,]
```

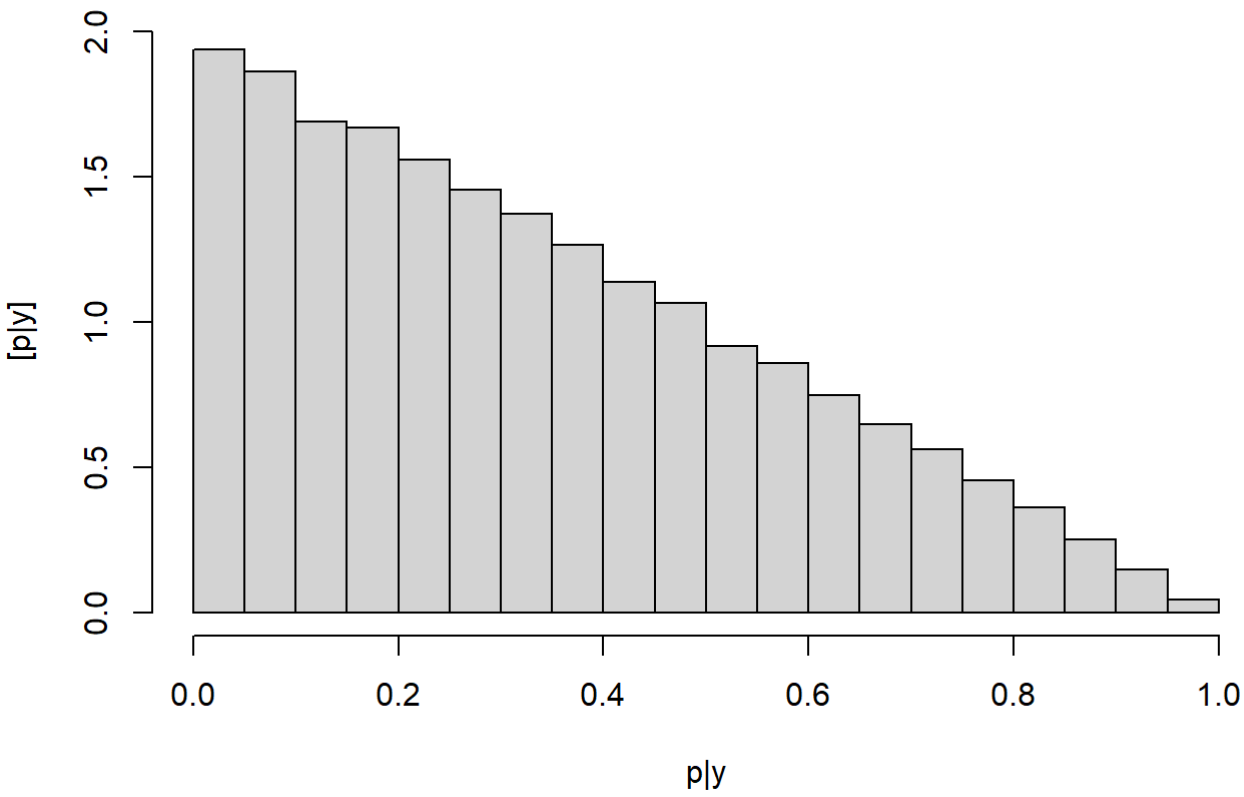
2.

```
hist(samples[,1],freq=FALSE,xlab="z|y",ylab="[z|y]",main="")
```



3.

```
hist(samples[,2],freq=FALSE,xlab="p|y",ylab="[p|y]",main="")
```



4.

```
mean(samples[,1]) # Expected value of Z_i
```

```
[1] 0.002
```

```
mean(samples[,2]) # Expected value of p
```

```
[1] 0.3341417
```

Looking at the $E(z_i|y_1, y_2)$ of z_i, the probability of z_i being a false negative in both tests is 0.002.

According to the expected value of $E(p|y_1, y_2)$, the prevalence rate of the diseases is 0.334. Given the distribution of $[p|y_1, y_2]$, it is helpful to determine the highest density interval for p, which determines that the prevalence rate should be in the range between 1.28e-5 and 0.779.

```
HDInterval::hdi(samples[,2], credMass = 0.95)
```

```
       lower        upper
1.283798e-05 7.778948e-01
attr(,"credMass")
[1] 0.95
```

5.

Find new prior

We performed a brief literature search on rabies prevalence in Kansas. Data for modelling the prior was retrieved by [The Kansas State University Rabies Laboratory] and made available trough [Kansas Environmental Public Health Tracking - KDHE].

The new prior can be written as follows:

$$[p] \sim Beta(\alpha_i, \beta_i) \; where \; \alpha_i = 2; \beta_i = 50$$

```
library(ggplot2)
library(tidyr)



p <- c(6/155, 4/166, 6/153, 6/153, 4/199, 3/234, 10/201, 2/198)

x =  seq(0, 1,0.001)

prior <- data.frame(
  # Old prior proposed on the excercise follows a Beta(1,1) or uniform distribution between 0
  old_prior = dbeta(x,1,1),
  # New prior approximates the rabies prevalence data retrieved from previous published litera
  new_prior = dbeta(x,2,80)
  ) %>%
  cbind(x) %>%
  pivot_longer(cols = c(old_prior, new_prior))




ggplot()+
  geom_histogram(aes(x = p, y = after_stat(density)),binwidth = 0.01, fill = NA, color = "blac
  geom_line(data = prior, aes(x = x, y = value, color = name), linewidth = 1)+
  guides(color = guide_legend(title = "Prior"))+
  scale_color_manual(values = c("green4", "red"), labels = c("New", "Old"))+
  theme_bw()
```

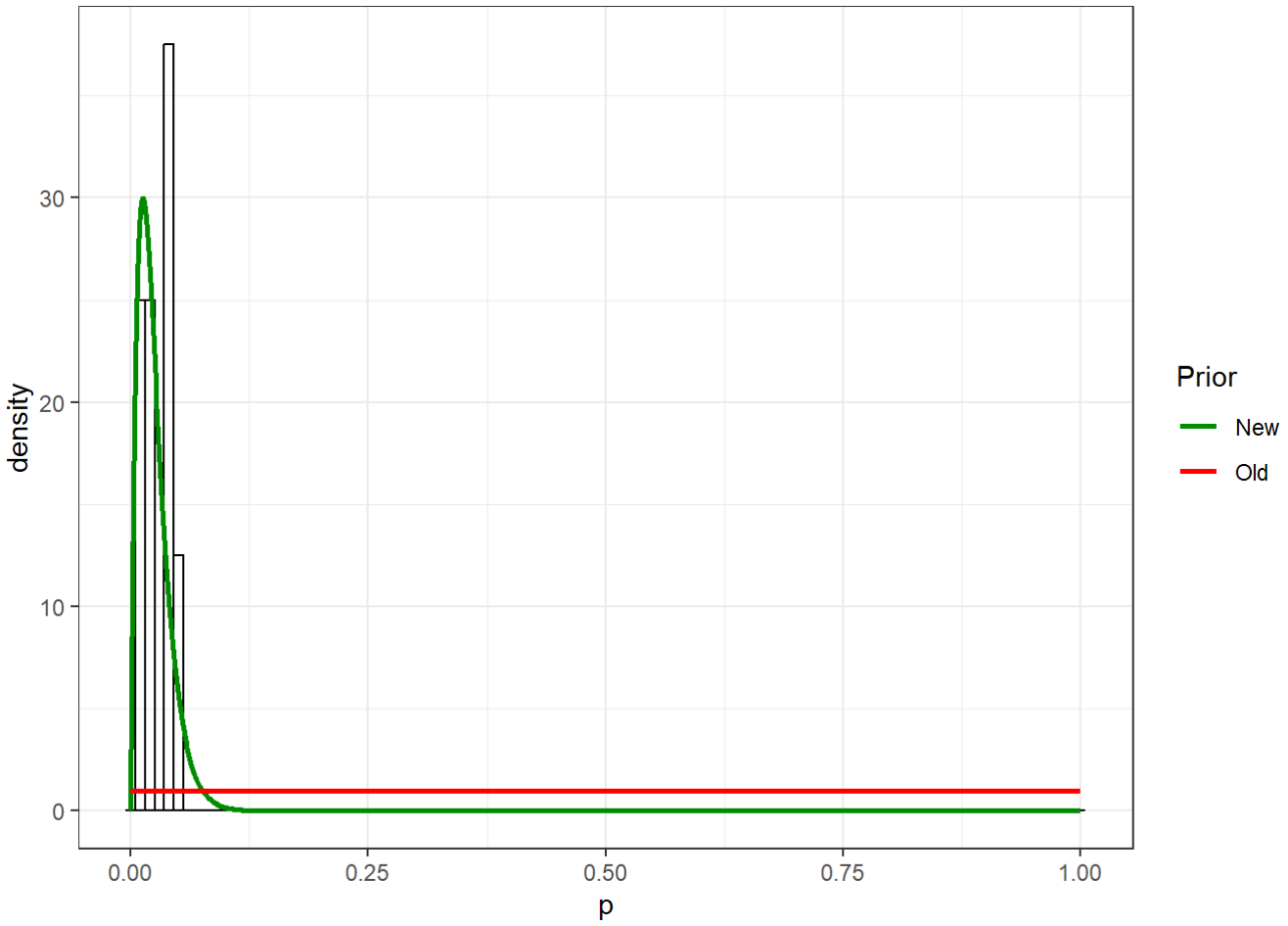


Figure. Probability density function of previously reported prevalence data in Kansas (bins), new prior (green) and old prior (red) distribution of rabies prevalence data.

6.

Fit bayessian model to data using new prior

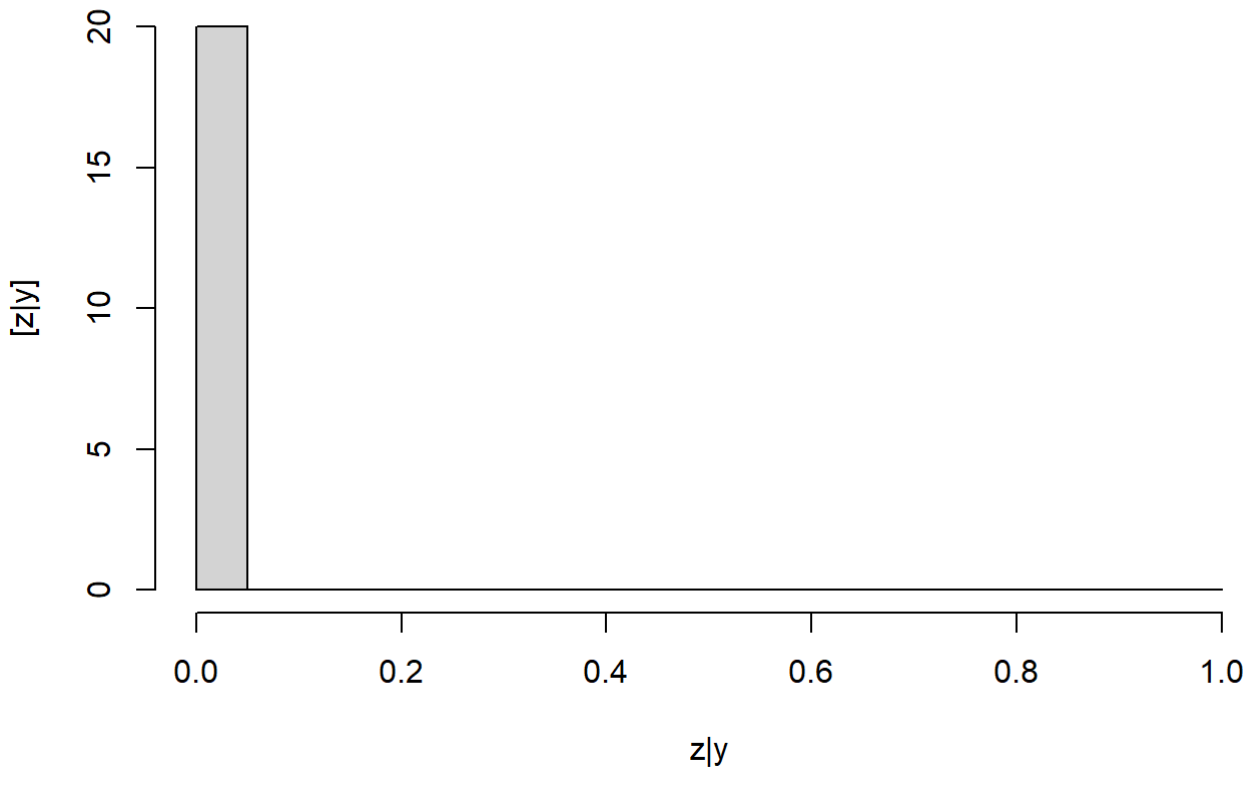
```
# Data
y1 <- 0 # Results from rapid test
y2 <- 0 # Results from PCR test


# Hyperparameters for priors
alpha.p <- 2
beta.p <- 80

# Fixed (assumed known) values of true positive rate
a1 <- 0.80
a2 <- 0.99


# Preliminary MCMC stuff
K <- 50000
samples <- matrix(,K,2)
colnames(samples) <- c("z1","p")
p.initial <- 0.01
```

```
# Gibbs sampler
for(k in 1:K){

  # Take one draw from the full conditional of z1 (i.e., [z1|p,y1,y2])
  p <- ifelse(k==1,p.initial,p)
  p.tilde <- (p*(1-a1)*(1-a2))/((p*(1-a1)*(1-a2))+1-p)
  z1 <- ifelse(min(c(y1,y2))==0,rbinom(1,1,p.tilde),1)

  # Take one draw from the full conditional of p (i.e., [p|z1,y1,y2])
  p <- rbeta(1,alpha.p+z1,beta.p+1-z1)

  # Save samples
  samples[k,] <- c(z1,p)
}

samples <- samples[10001:K,]
```

7.

Histogram representation of the posterior distribution of the true rabies status using the new prior.

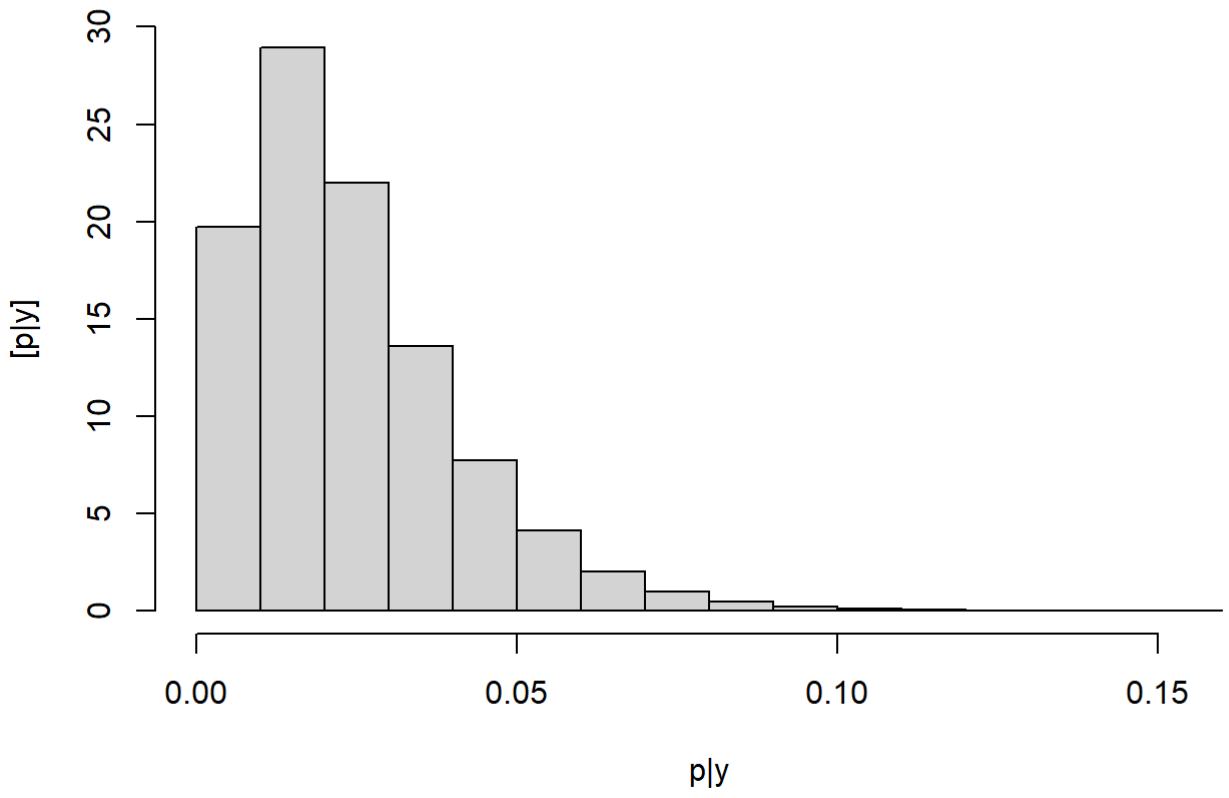
```
# Histogram representation of the posterior distributions
hist(samples[,1],freq=FALSE,xlab="z|y",ylab="[z|y]",main="")
```



8.

Posterior distribution of the prevalence rate parameter using the new prior.

```
hist(samples[,2],freq=FALSE,xlab="p|y",ylab="[p|y]",main="")
```



9.

Using the new prior, the probability of the bat analyzed having rabies (being a false negative in both test) is close to zero looking at the expected value of $z_i|y_1, y_2$

```
# True rabies status posterior summaries
mean(samples[,1])
```

```
[1] 5e-05
```

From our posterior distribution outputs our expected value for the prevalence rate $E(p|y_1, y_2)$ is 0.02423536. The highest density interval, which accounts for the 95% of the values, is located between 0.0004512952 and 0.0568963900.

```
# Prevalence rate posterior summaries
mean(samples[,2])
```

```
[1] 0.02423536
```

```r
HDInterval::hdi(samples[,2], credMass = 0.95)
```

```
       lower        upper
0.0004512952 0.0568963900
attr(,"credMass")
[1] 0.95
```

Compared to the results with the old prior distribution, the new posterior inform a lower possibility of both tests being a false negative. Also the prevalence rate conditioned on the test result has a lower expected value than the obtained with the old prior, while the HDI has been narrowed compared to the previous results.