hw4 for stat341

Zhihong Zhang

Feb 13th,2017

Q: 3H1–3H5,4E1-4E2 4M1 4M2

Example:

```
# load the birth1 and birth2 data vectors
data(homeworkch3, package = "rethinking")
# put them into a data frame
Birth <- data_frame(
first = birth1,
second = birth2
)
# tally up the counts
tally( ~ first + second, data = Birth, margins = TRUE)
```

```
##        second
## first     0   1 Total
##   0      10  39    49
##   1      30  21    51
##   Total  40  60   100
```

```
# another way to summarize:
Birth %>%
# group by family type
group_by(first, second) %>%
summarise(
# how many families of this type
families = n(),
# total boys in such families
boys = sum(first + second),
# total girls in such families
girls = sum(2 - first - second)
)
```

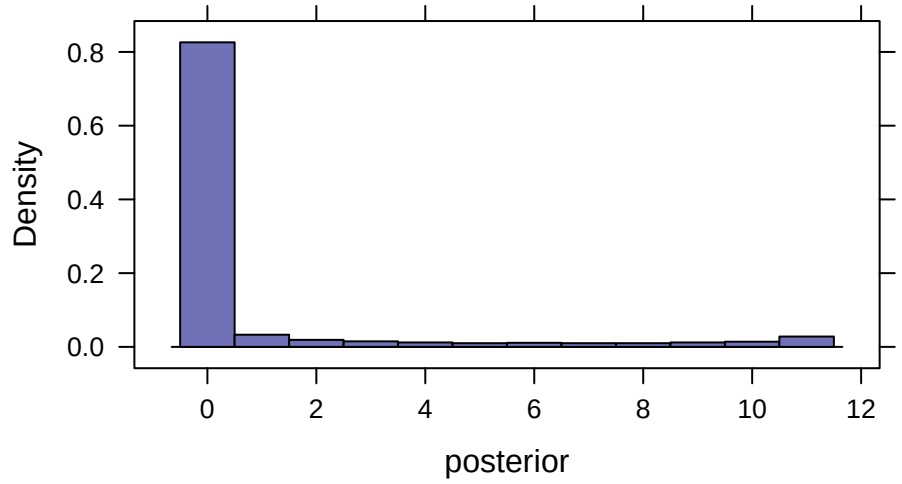
```
## Source: local data frame [4 x 5]
## Groups: first [?]
##
##   first second families  boys girls
##   <dbl>  <dbl>    <int> <dbl> <dbl>
## 1     0      0       10     0    20
## 2     0      1       39    39    39
## 3     1      0       30    30    30
## 4     1      1       21    42     0
```

3H1. Using grid approximation, compute the posterior distribution for the probability of a birth being a boy. Assume a uniform prior probability. Which parameter value maximizes the posterior probability? p parameter maximize the posterior probability *Solution:*

```
Grid <-
  expand.grid(p = seq(0, 1, length.out = 1000)) %>%       # create grid of values for p
  mutate(                                                 # add additional variables
```

```
    prior =dunif(p,0,1),                                  # uniform prior, value gets recycled
    likelihood = dbinom(111, size = 200, prob = p),              #  probabilty
    posterior_raw = prior * likelihood,                        # kernel of posterior
    posterior1 = posterior_raw / sum(posterior_raw),          # easy normalization
    posterior = posterior_raw / sum(posterior_raw) / 0.001 # fancy normalization
  )

histogram(p~posterior,Grid,width=1)
```



3H2. Using the sample function, draw 10,000 random parameter values from the posterior distribution you calculated above. Use these samples to estimate the 50%, 89%, and 97% highest posterior density intervals.

Solution:

```
BinomGrid <-
  expand.grid(p = seq(0, 1, by = 0.0001)) %>%       # create grid of values for p
  mutate(                                             # add additional variables
    prior =dunif(p,0,1),                              # uniform prior, value gets recycled
    likelihood = dbinom(111, size = 200, prob = p),              # binomial probabilty
    posterior_raw = prior * likelihood,                        # kernel of posterior
    posterior1 = posterior_raw / sum(posterior_raw),          # easy normalization
    posterior = posterior_raw / sum(posterior_raw) / 0.001 # fancy normalization
  )

Boys.Post <- sample(BinomGrid,1e4,replace=TRUE,prob=BinomGrid$posterior)
PI(Boys.Post$p,prob=0.5)
```

```
##    25%    75%
## 0.5307 0.5774
```

```
PI(Boys.Post$p,prob=0.89)
```

```
##     5%    94%
## 0.4990 0.6105
```

```
PI(Boys.Post$p,prob=0.97)
```

```
##       2%      98%
```

```
## 0.478497 0.629703
```

```
#PI(posterior, prob = 0.5)
#PI(posterior, prob = 0.89)
#PI(posterior, prob = 0.97)
```

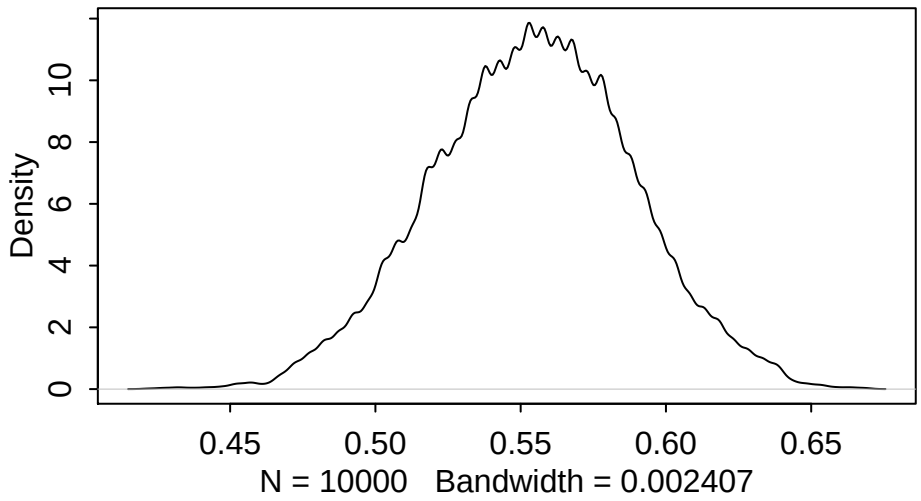
3H3. Use rbinom to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births). There are many good ways to visualize the simulations, but the dens command (part of the rethinking package) is probably the easiest way in this case. Does it look like the model fits the data well? That is, does the distribution of predictions include the actual observation as a central, likely outcome?

Solution:

```
# define grid
p_grid <- seq( from=0 , to=1 , length.out=200 )

# define prior
prior <- rep( 1 , 200 )
likelihood = dbinom(111, size =  200, prob =p_grid )
posterior = prior * likelihood
posterior_raw = prior * likelihood                    # kernel of posterior
posterior = posterior_raw / sum(posterior_raw)        # easy normalization
posterior2 = sample(p_grid, posterior, size = 1e4, replace = TRUE)

dens(posterior2)
```



3H4. Now compare 10,000 counts of boys from 100 simulated first borns only to the number of boys in the first births, birth1. How does the model look in this light?

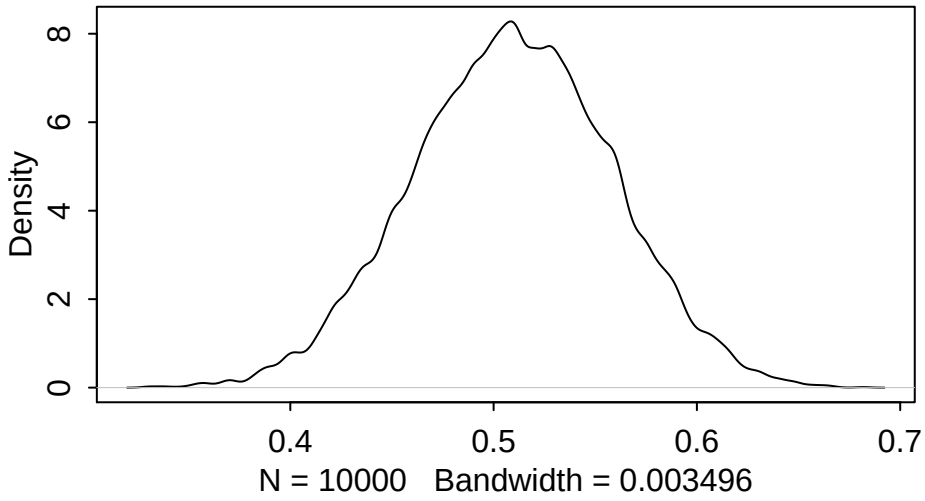
Solution:

```
sum(birth1)     # number of the boy born in first birth
```

```
## [1] 51
```

```
# define grid
p_grid <- seq( from=0 , to=1 , length.out=10000 )
```

```
# define prior
prior <- rep( 1 , 10000 )
likelihood = dbinom(51, size =  100, prob =p_grid )  #read from example, there is 51 boys born in first
posterior = prior * likelihood
posterior_raw = prior * likelihood                    # kernel of posterior
posterior = posterior_raw / sum(posterior_raw)        # easy normalization
posterior2 = sample(p_grid, posterior, size = 1e4, replace = TRUE)
dens(posterior2)
```



3H5. The model assumes that sex of first and second births are independent. To check this assumption, focus now on second births that followed female first borns. Compare 10,000 simulated counts of boys to only those second births that followed girls. To do this correctly, you need to count the number of first borns who were girls and simulate that many births, 10,000 times. Compare the counts of boys in your simulations to the actual observed count of boys following girls. How does the model look in this light? Any guesses what is going on in these data?

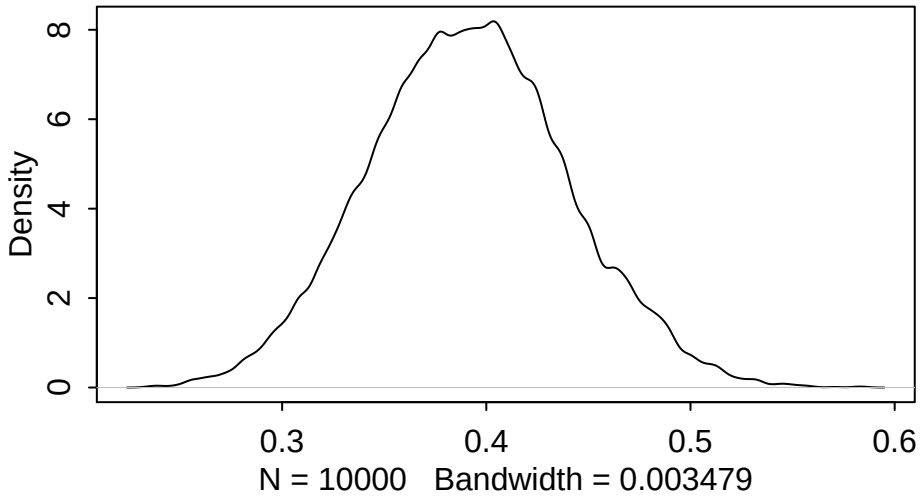
Solution:

```
sum(birth1)     # number of the boy born in first birth
```

```
## [1] 51
```

```
# define grid
p_grid <- seq( from=0 , to=1 , length.out=10000 )

# define prior
prior <- rep( 1 , 10000 )
likelihood = dbinom(39, size =  100, prob =p_grid )
posterior = prior * likelihood
posterior_raw = prior * likelihood                    # kernel of posterior
posterior = posterior_raw / sum(posterior_raw)        # easy normalization
posterior2 = sample(p_grid, posterior, size = 1e4, replace = TRUE)
dens(posterior2)
```



The guess is that the indepence of the data may still need to be checked. The family is from the people who prefered more on the boy, and it may cause the data to be biased.

4E1. In the model definition below, which line is the likelihood? $y_i \sim Normal(\mu, \sigma)$
$\mu \sim Normal(0, 10)$
$\sigma \sim Uniform(0, 10)$

Solution:

$y_i \sim Normal(\mu)$ is the likelihood.

4E2. In the model definition just above, how many parameters are in the posterior distribution?

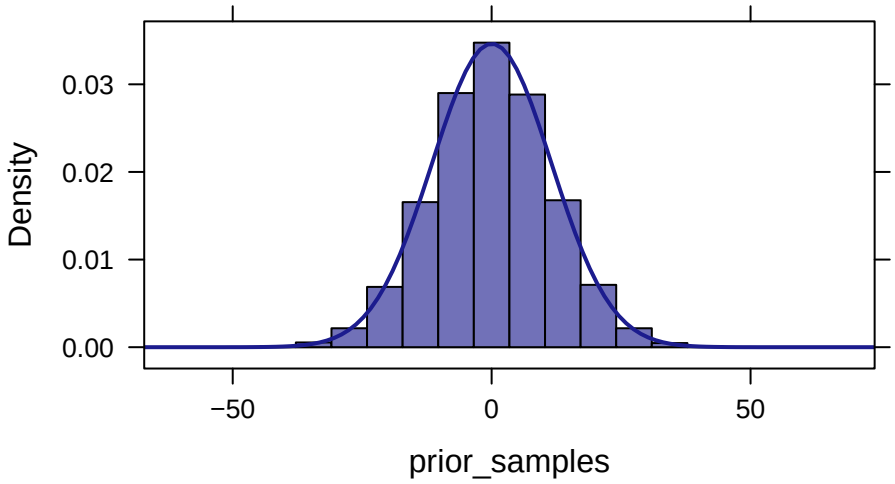
Solution:

There are 2 parameters in the posterior distribution. It is μ *and* σ

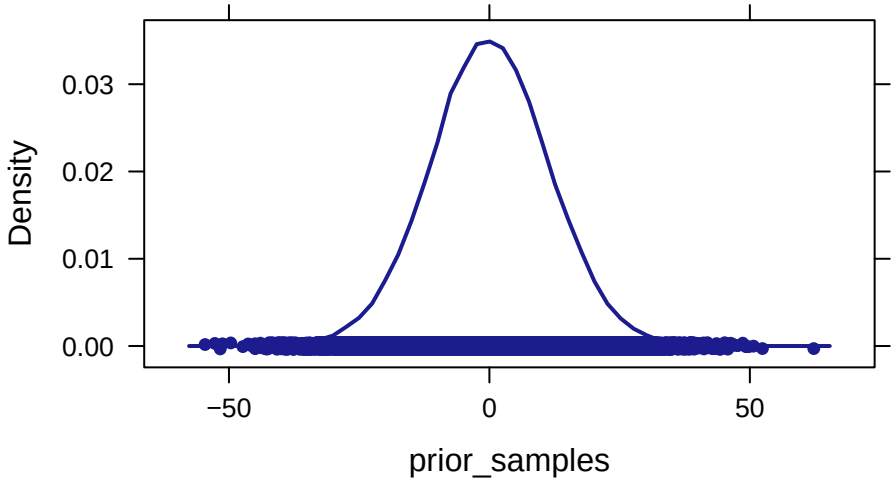
4M1. For the model definition below, simulate observed heights from the prior (not the posterior).
$y_i \sim Normal(\mu, \sigma)$
$\mu \sim Normal(0, 10)$
$\sigma \sim Uniform(0, 10)$

Solution:

```
# choose 100000 run.
mu = rnorm(1e5, 0, 10)
sigma = runif(1e5, 0,10)
prior_samples <- rnorm(1e5,mu,sigma)
histogram(~prior_samples, fit = "normal")
```




```
densityplot(~prior_samples)
```



4M2. Translate the model just above into a map formula.

Solution:

```
f_list <- alist(
  height   ~ dnorm( mu, sigma ),
  mu  ~ dnorm( 0,  10 ),
  sigma ~ dunif( 0,  10 )
)
#Then use map(f_list,data) to make a plot if we have data
```