hw1

GISH

2019撟3<9c><88>1

## Question 1

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?

library(tidyverse)

## -- Attaching packages ------------------------------------------------------ tidyverse 1.2.1 --

## √ ggplot2 3.1.0 √ purrr 0.3.0   
## √ tibble 2.0.1 √ dplyr 0.8.0.1  
## √ tidyr 0.8.2 √ stringr 1.4.0   
## √ readr 1.3.1 √ forcats 0.4.0

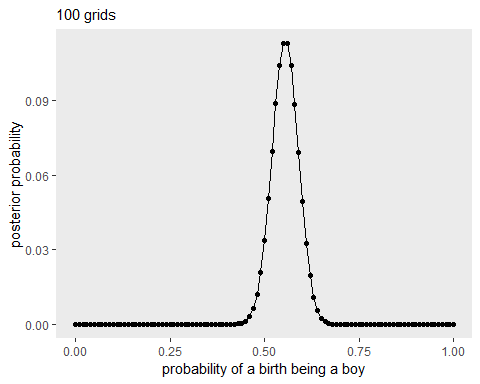
## -- Conflicts --------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

birth1 <- c(1,0,0,0,1,1,0,1,0,1,0,0,1,1,0,1,1,0,0,0,1,0,0,0,1, 0,0,0,0,1,1,1,0,1,0,1,1,1,0,1,0,1,1,0,1,0,0,1,1,0, 1,0,0,0,0,0,0,0,1,1,0,1,0,0,1,0,0,0,1,0,0,1,1,1,1, 0,1,0,1,1,1,1,1,0,0,1,0,1,1,0,1,0,1,1,1,0,1,1,1,1)  
  
birth2 <- c(0,1,0,1,0,1,1,1,0,0,1,1,1,1,1,0,0,1,1,1,0,0,1,1,1, 0,1,1,1,0,1,1,1,0,1,0,0,1,1,1,1,0,0,1,0,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,1,0,1,1,1,0,0,0,0, 0,0,1,0,0,0,1,1,0,0,1,0,0,1,1,0,0,0,1,1,1,0,0,0,0)   
  
family\_birth\_data <- tibble(birth1,birth2)  
  
birth\_grid\_posterior <- tibble(p\_grid = seq(from = 0, to = 1, length.out = 101), #100 grids  
prior = 1) %>%   
mutate(likelihood = dbinom(sum(birth1) + sum(birth2), size = 200, prob = p\_grid)) %>%  
# binomial distribution with boys as sucess trials  
mutate(unstandard\_post = likelihood \* prior) %>%   
mutate(posterior = unstandard\_post / sum(unstandard\_post))  
birth\_grid\_posterior # show posterior dataframe

## # A tibble: 101 x 5  
## p\_grid prior likelihood unstandard\_post posterior  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 0. 0. 0.   
## 2 0.01 1 1.11e-164 1.11e-164 2.23e-164  
## 3 0.02 1 1.17e-131 1.17e-131 2.34e-131  
## 4 0.03 1 1.64e-112 1.64e-112 3.31e-112  
## 5 0.04 1 4.83e- 99 4.83e- 99 9.70e- 99  
## 6 0.05 1 1.09e- 88 1.09e- 88 2.18e- 88  
## 7 0.06 1 2.61e- 80 2.61e- 80 5.24e- 80  
## 8 0.07 1 2.72e- 73 2.72e- 73 5.46e- 73  
## 9 0.08 1 2.84e- 67 2.84e- 67 5.71e- 67  
## 10 0.09 1 5.11e- 62 5.11e- 62 1.03e- 61  
## # ... with 91 more rows

## Question 1 Plot

#plotting  
birth\_grid\_posterior %>%  
ggplot(aes(x = p\_grid, y = posterior)) +  
geom\_point() +  
geom\_line() +   
labs(subtitle = "100 grids",  
x = "probability of a birth being a boy",  
y = "posterior probability") +  
theme(panel.grid = element\_blank())



# Which parameter value maximizes the posterior probability?

birth\_grid\_posterior %>% summarise(Max\_Vals\_post = max(posterior), p\_grid = which.max(posterior))

## # A tibble: 1 x 2  
## Max\_Vals\_post p\_grid  
## <dbl> <int>  
## 1 0.113 56

So the parameter is 56th p\_grid, i.e. when probability of a birth being a boy =0.56 , it has highest posterior probability ~= 0.113

## Question 2

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. (hint: use mode\_hdi() in tidybayes packages)

samples <- sample(birth\_grid\_posterior$p\_grid, #select p\_grid (parameter values)  
prob = birth\_grid\_posterior$posterior, # posterior probability of each grid  
size = 1e4, # 10000 samples  
replace = T) # with replacement  
head(samples,20) #show first 20 samples

## [1] 0.53 0.53 0.57 0.58 0.55 0.57 0.54 0.51 0.56 0.53 0.58 0.58 0.55 0.56  
## [15] 0.53 0.57 0.59 0.58 0.56 0.53

samples <- tibble(samples = sample(birth\_grid\_posterior$p\_grid, prob = birth\_grid\_posterior$posterior, size = 1e4, replace = T)) # sample tibblized  
glimpse(samples)

## Observations: 10,000  
## Variables: 1  
## $ samples <dbl> 0.57, 0.55, 0.54, 0.58, 0.53, 0.58, 0.53, 0.62, 0.56, ...

library(tidybayes)

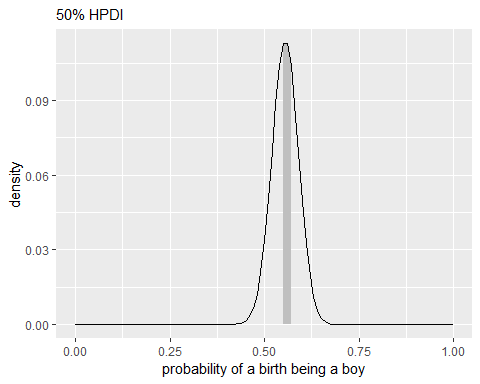
## NOTE: As of tidybayes version 1.0, several functions, arguments, and output column names  
## have undergone significant name changes in order to adopt a unified naming scheme.  
## See help('tidybayes-deprecated') for more information.

mode\_hdi(samples$samples, .width = c(.5, .89, .97))

## y ymin ymax .width .point .interval  
## 1 0.5503033 0.54 0.58 0.50 mode hdi  
## 2 0.5503033 0.49 0.60 0.89 mode hdi  
## 3 0.5503033 0.47 0.62 0.97 mode hdi

## Question 2 Plot (50% hdpi)

birth\_grid\_posterior %>%  
ggplot(aes(x = p\_grid)) +  
geom\_ribbon(data = birth\_grid\_posterior %>% filter(p\_grid > mode\_hdi(samples$samples, .width  
= .5)[, "ymin"] &  
p\_grid < mode\_hdi(samples$samples, .width = .5)[, "ymax"]),  
aes(ymin = 0, ymax = posterior),  
fill = "grey75") +  
geom\_line(aes(y = posterior)) +  
labs(subtitle = "50% HPDI",  
x = "probability of a birth being a boy",  
y = "density")



## Question 3

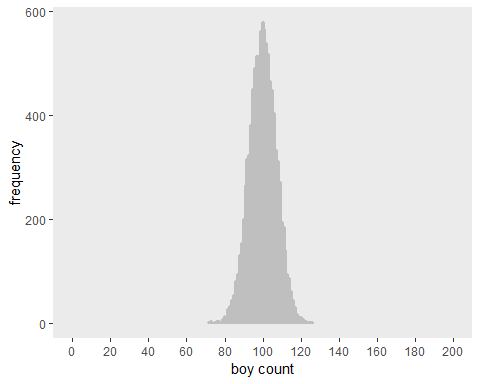
Use rbinom function to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data(111 boys out of 200 births). There are many good ways to visualize 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?

set.seed(123)  
d <- tibble(number\_of\_boys = rbinom(10000, size = 200, prob = .5))  
d %>%  
group\_by(number\_of\_boys) %>% count() %>% mutate(proportion = n / nrow(d))

## # A tibble: 54 x 3  
## # Groups: number\_of\_boys [54]  
## number\_of\_boys n proportion  
## <int> <int> <dbl>  
## 1 72 1 0.0001   
## 2 73 3 0.000300  
## 3 75 2 0.0002   
## 4 76 3 0.000300  
## 5 77 4 0.0004   
## 6 78 2 0.0002   
## 7 79 7 0.0007   
## 8 80 14 0.0014   
## 9 81 11 0.0011   
## 10 82 26 0.0026   
## # ... with 44 more rows

## Question 3 Plot

d %>%  
ggplot(aes(x = number\_of\_boys)) +   
geom\_histogram(binwidth = 1,   
center = 0,  
color = "grey75",   
size = 1) +   
scale\_x\_continuous("boy count",   
breaks = seq(from = 0, to = 200, by = 20)) +   
ylab("frequency") +  
coord\_cartesian(xlim = 0:200) +  
theme(panel.grid = element\_blank())



Compared to question 1 plot, it looks like the model fits the data well. Making the distribution of predictions include the acutal observation as a central likely outcome.

## Question 4

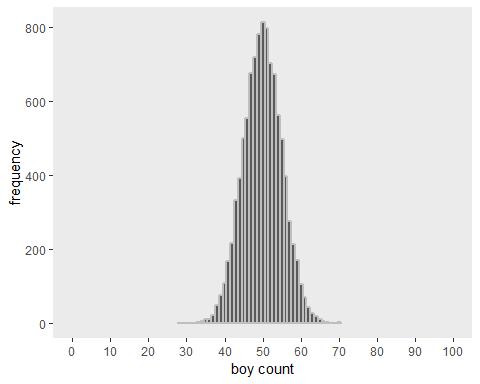
Now compare 10,000 counts of boys from 100 simulated ﬁrst borns only to the number of boys in the ﬁrst births, birth1. How does the model look in this light?

## Question 4 Plot A: Simulation Plot

d <- tibble(number\_of\_boys = rbinom(10000, size = 100, prob = .5))  
d %>%  
group\_by(number\_of\_boys) %>% count() %>% mutate(proportion = n / nrow(d))

## # A tibble: 40 x 3  
## # Groups: number\_of\_boys [40]  
## number\_of\_boys n proportion  
## <int> <int> <dbl>  
## 1 28 1 0.0001  
## 2 30 1 0.0001  
## 3 31 1 0.0001  
## 4 32 1 0.0001  
## 5 33 2 0.0002  
## 6 34 4 0.0004  
## 7 35 11 0.0011  
## 8 36 11 0.0011  
## 9 37 21 0.0021  
## 10 38 49 0.0049  
## # ... with 30 more rows

d %>%  
ggplot(aes(x = number\_of\_boys)) +   
geom\_histogram(binwidth = 1,   
center = 0,  
color = "grey75",   
size = 1) +   
scale\_x\_continuous("boy count",   
breaks = seq(from = 0, to = 100, by = 10)) +   
ylab("frequency") +  
coord\_cartesian(xlim = 0:100) +  
theme(panel.grid = element\_blank())

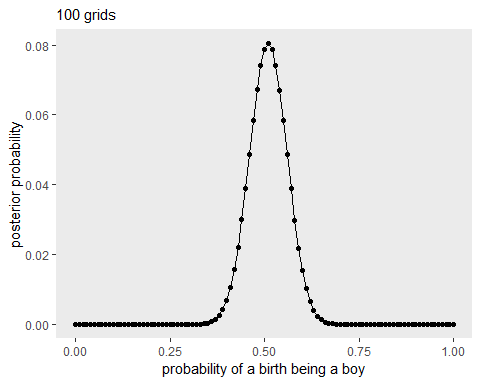


## Question 4 Plot B: Actual Observations Plot

birth\_grid\_posteriorB <- tibble(p\_grid = seq(from = 0, to = 1, length.out = 101), #100 grids  
prior = 1) %>%   
mutate(likelihood = dbinom(sum(birth1), size = 100, prob = p\_grid)) %>%  
# binomial distribution with boys as sucess trials  
mutate(unstandard\_post = likelihood \* prior) %>%   
mutate(posterior = unstandard\_post / sum(unstandard\_post))  
birth\_grid\_posteriorB # show posterior dataframe

## # A tibble: 101 x 5  
## p\_grid prior likelihood unstandard\_post posterior  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0 1 0. 0. 0.   
## 2 0.01 1 6.04e-74 6.04e-74 6.11e-74  
## 3 0.02 1 8.28e-59 8.28e-59 8.36e-59  
## 4 0.03 1 4.79e-50 4.79e-50 4.84e-50  
## 5 0.04 1 6.79e-44 6.79e-44 6.85e-44  
## 6 0.05 1 3.56e-39 3.56e-39 3.59e-39  
## 7 0.06 1 2.31e-35 2.31e-35 2.34e-35  
## 8 0.07 1 3.56e-32 3.56e-32 3.59e-32  
## 9 0.08 1 1.90e-29 1.90e-29 1.92e-29  
## 10 0.09 1 4.51e-27 4.51e-27 4.56e-27  
## # ... with 91 more rows

birth\_grid\_posteriorB %>%  
ggplot(aes(x = p\_grid, y = posterior)) +  
geom\_point() +  
geom\_line() +   
labs(subtitle = "100 grids",  
x = "probability of a birth being a boy",  
y = "posterior probability") +  
theme(panel.grid = element\_blank())



sum(birth1)

## [1] 51

sum(birth2)

## [1] 60

Because birth1’s data are more evenly distributed(51 out of 100 childs are boys, p1 = 0.51) than birth2’s(60 of 100 p2=0.6), simulated data(with p =0.5) are acutally more similar to birth1, as forementioned plot showed