Methods 4 – Portfolio 1

- *Type:* Group assignment
- *Due:* 6 March 2022, 23:59

Okay here is a re-skinned version of some of McElreath's Exercises.

Have fun :)

Trigger alert for anyone who has recently experienced a pandemic.

– *Peter and Chris*

Pandemic Exercises

1) Testing Efficiency

Imagine there was a global pandemic.

It's a bit difficult, I know.

Maybe a new version of the old SARS-CoV turns out to be really infectious, or something like that.

A test is developed that is cheap and quick to use, and the government asks you to determine its efficiency.

To do this, they find X people that they know for sure are infected, and X people that they know for sure are not infected. *NB: This is not always possible. For example, there is an ongoing global pandemic in the real world - maybe you heard of it -where a 100% sure test doesn't exist, as far as I know. But let's ignore that. The government finds a wizard who can tell for sure, but he wants a lot of money and he's really slow too.*

Okay, so X infected people take the test, and X uninfected people take the test. See the results below. P means positive, N means negative.

- Infected:

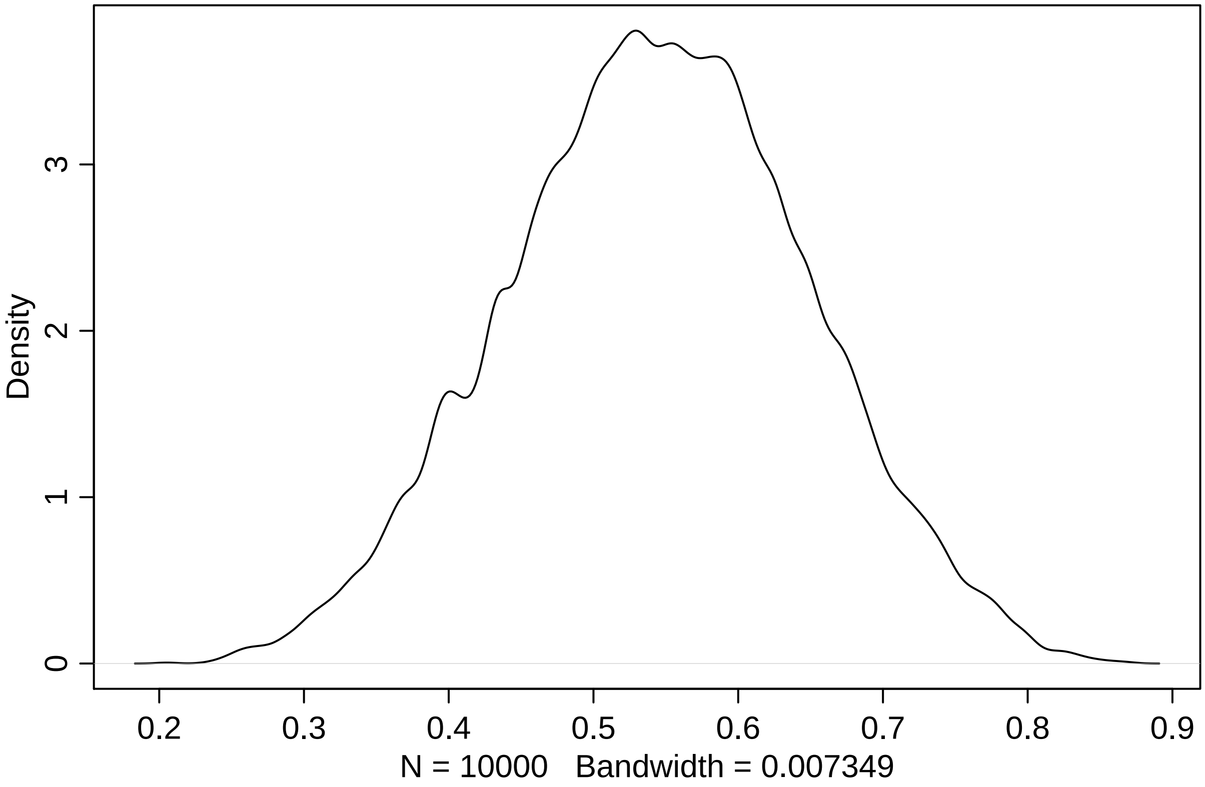
$$P, N, P, P, N, P, P, N, N, N, P, P, N, P, P, N, N, P, N, P$$

- Uninfected:

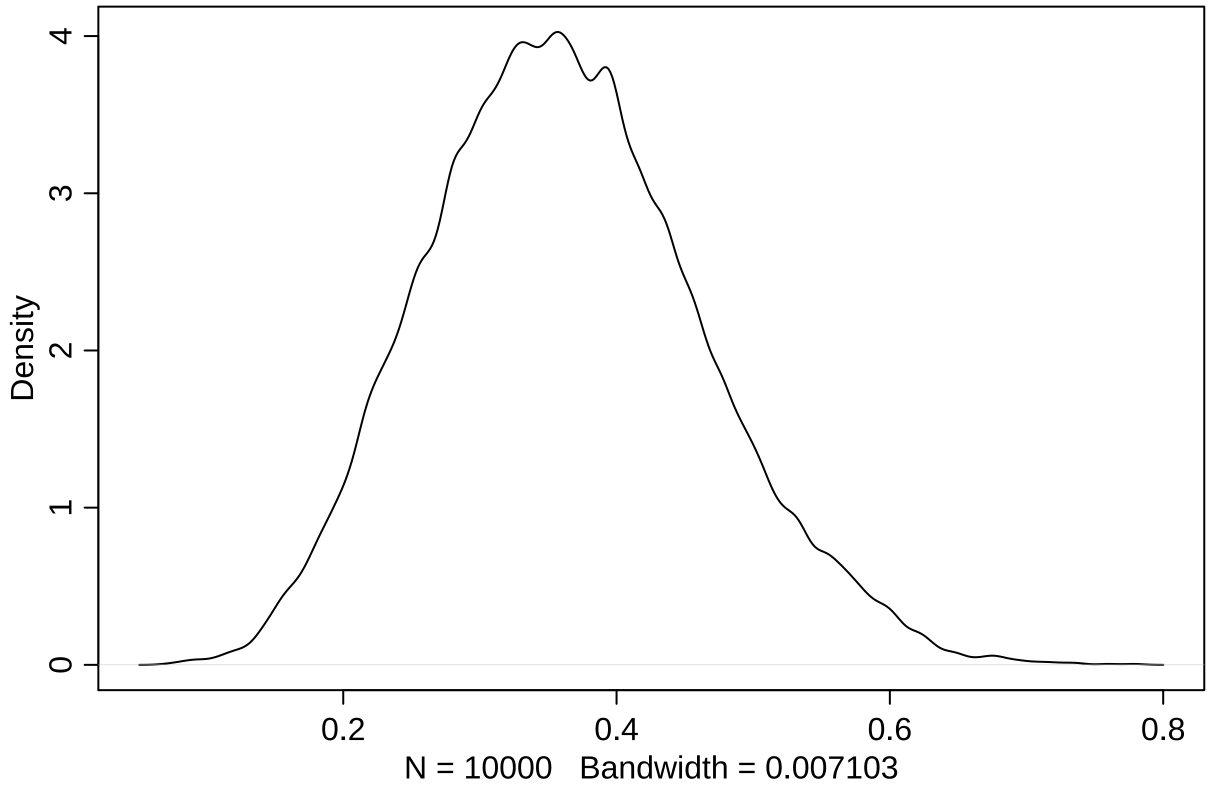
$$P, N, N, P, N, P, P, N, N, N, P, N, N, N, N, P, P, N, N, N$$

A) Estimate the probabilities of testing positive given that you're infected, and given that you're not infected. Use the grid approximation method as in the book. Use a prior you can defend using. Report the full posterior probability distribution for each case (we can do better than just a single value!).

```r
# Infected:
p_grid <- seq(from=0 , to=1 , length.out=1000 )
prior <- rep(1 , 1000 ) # a flat prior as we assume it's equally likely to be infected (it's a very wic
likelihood <- dbinom(11 , size=20 , prob=p_grid) # 11 out of 20
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)

set.seed(100)
samples_infected <- sample(p_grid , prob=posterior , size=1e4 , replace=TRUE)
dens(samples_infected)
```



```r
# Uninfected:
p_grid <- seq(from=0 , to=1 , length.out=1000 )
prior <- rep(1 , 1000 ) # a flat prior as we assume it's equally likely to be infected (it's a very wic
likelihood <- dbinom(7 , size=20 , prob=p_grid) # 7 out of 20
posterior <- likelihood * prior
posterior <- posterior / sum(posterior)

set.seed(100)
samples_uninfected <- sample(p_grid , prob=posterior , size=1e4 , replace=TRUE)
dens(samples_uninfected)
```



B) The government says that they find probability distributions difficult to use. They ask you to provide them with a confidence interval of 95% within which the 'real' probability can be found. Do it.

```
HPDI(samples_infected, prob = 0.95) # 0.35 - 0.75
```

```
##     |0.95     0.95|
## 0.3483483 0.7497497
```

```
HPDI(samples_uninfected, prob = 0.95) # 0.17 - 0.56
```

```
##     |0.95     0.95|
## 0.1711712 0.5585586
```

C) The government says that their voters find confidence intervals difficult to read. In addition, they are so wide that it looks like the government doesn't know what they're doing. They want a point estimate instead. Give them one.

```
chainmode(samples_infected) # 0.53
```

```
## [1] 0.53483
```

```
chainmode(samples_uninfected) # 0.35
```

```
## [1] 0.350261
```

Conclusion: Summary of posterior predictive distributions:

For people infected:

By examining the two posterior probability distributions, we get the impression that there is a higher probability of testing positive if you are actually infected than if you are not infected (luckily enough). The confidence interval (HDPI interval) tells us that there is a 95% chance that the true probability of testing positive given that you are infected is between 35% and 75%. This interval is pretty wide, so therefor we also looked at the point estimate, which is the maximum value of the posterior predictive distribution (the mode). Using the point estimate, we can (carefully) conclude that there's around a 53% probability of being infected given a positive test.

For people NOT infected:

For the people not infected, the HDPI interval tells us that there is a 95% chance that the true probability of testing positive is between 17% and 56%. According to the point estimate, there is more precisely a 35% probability of testing positive given that you are not infected. The point estimate is useful, because it is easy to interpret, but nevertheless it is less informative than reporting a full distribution.
We can indeed conclude that there is a higher probability of getting a positive test result, given that you are actually infected.

2) Dark Cellars

Months pass. Thousands of people are tested by the wizards of the world governments. A fancy company analyses the data, and determine, with very high confidence they say, the probability of testing positive with the current test. They give the following point estimates:

- A 53% chance of testing positive if you are infected.
- A 45% chance of testing positive if you are not infected.

NB: These numbers also happen to be real estimates for the efficiency of the COVID kviktest[^1]. Remember that the actual Danish government doesn't have any wizards, though.

A) You are sitting in your dark cellar room, trying to come up with an apology to the Danish government, when you receive a positive test result on your phone. Oh, that party last weekend. In order to fight the boredom of isolation life, you start doing statistical inference. Estimate the probability that you are infected, given that it is *a priori* equally likely to be infected or not to be.

```
Pr_pos_inf <- 0.53 # probability of getting a positive test given you're infected
Pr_pos_uninf <- 0.45 # probability of getting a positive test given you're NOT infected
Pr_inf <- 0.5 # general probability in population (prior)

Pr_pos <- Pr_pos_inf * Pr_inf + Pr_pos_uninf * (1 - Pr_inf) # general prob of getting a positive result

# the probability of being infected given a positive test (this is what we want)
Pr_inf_pos <- Pr_pos_inf*Pr_inf / Pr_pos
Pr_inf_pos # 0.54
```

```
## [1] 0.5408163
```

There's a 54% probability of being infected given a positive test result

B) A quick Google search tells you that about 546.000[^2] people in Denmark are infected right now. Use this for a prior instead.

```
Pr_inf_new <- 546000/5.8e6
Pr_inf_new # 0.0941 (9% of people are infected)
```

```
## [1] 0.09413793
```

```
Pr_pos <- Pr_pos_inf * Pr_inf_new + Pr_pos_uninf * (1 - Pr_inf_new)

Pr_inf_pos <- Pr_pos_inf*Pr_inf_new / Pr_pos
Pr_inf_pos
```

```
## [1] 0.1090486
```

After updating our model, so that the prior is informed with the *actual* amount of infected people in the population: We see there is only 11% probability of actually being infected given that you get a positive test result.

C) A friend calls and says that they have been determined by a wizard to be infected. You and your friend danced tango together at the party last weekend. It has been estimated that dancing tango with an infected person leads to an infection 32% of the time[^3]. Incorporate this information in your estimate of your probability of being infected.

```
Pr_inf_tango <- 0.32

Pr_pos <- Pr_pos_inf * Pr_inf_tango + Pr_pos_uninf * (1 - Pr_inf_tango)

Pr_inf_pos_tango <- Pr_pos_inf*Pr_inf_tango / Pr_pos
Pr_inf_pos_tango # 0.3566
```

```
## [1] 0.3566022
```

We assume that we disregard the information that only 9% of the population is actually infected. We then apply the 32% prob of being infected given tango dance as our new prior. This gives a 35.66% probability of being infected given a positive test and a tango dance.

D) You quickly run and get two more tests. One is negative, the other positive. Update your estimate.

- A 53% chance of testing positive if you are infected.
- A 45% chance of testing positive if you are not infected.

```
# taking our previous posterior making it a prior, but IN ODDS
Odds_inf_pos_tango <- Pr_inf_pos_tango/(1-Pr_inf_pos_tango) # turning the prior in probabilities into a

# Bayes factor given a positive test result
Bayes_fact_pos <- 0.53/0.45

# Bayes factor given a negative test result
Bayes_fact_neg <- (1-0.53)/(1-0.45)

# now we can multiply the prior in odds with Bayes factor, first given a positive result and then multi
Updated_odds <- Odds_inf_pos_tango*Bayes_fact_pos*Bayes_fact_neg

# turning back into probabilities
Pr_inf_pos_tango_tests <- Updated_odds/(1+Updated_odds)
Pr_inf_pos_tango_tests
```

```
## [1] 0.358082
```

Given another positive test result and another negative test result the probability of being infected is now 35.80%. This is slightly above over estimate from before, but it seems as if the two tests cancel each other.

E) In a questionnaire someone sent out for their exam project, you have to answer if you think you are infected. You can only answer yes or no (a bit like making a point estimate). What do you answer?

Answer (E):

In this calculation we've worked with the calculations in *odds*. By exploiting that updating can be done by multiplying a prior with bayes factor. We assume that there's a 53% chance of testing positive if you are infected and a 45% chance of testing positive if you are not infected.

Assuming we've been dancing tango with an infected individual and have gotten 1 positive and 1 negative test, we calculate that there's only 35.8% probability of actually being infected. Therefore we say NO, we don't not believe that we are infected.

F) You are invited to a party. They ask if you are infected. They also say that they would prefer if you used an asymmetric loss function when making your decision: it is three times worse to falsely answer not infected, than to falsely answer infected. What do you answer?

```r
prob_inf <- Pr_inf_pos_tango_tests
prob_uninf <- 1-Pr_inf_pos_tango_tests

prob_inf*3 # cost of being infected
```

```
## [1] 1.074246
```

```r
prob_uninf*1 # cost of not being infected
```

```
## [1] 0.641918
```

In part "D" we calculated the probability of being infected after dancing tango and having the two test, this is "Pr_inf_pos_tango_tests". The probability of being infected is deemeed 3 times as bad as not being infected, hence comparing the loss in the two situations gives us a relative measure of cost.

As the cost associated with actually being infected after the two test is larger than the cost of not being infected, we decide to stay at home.

3) Causal Models

A problem from our textbook *Statistical Rethinking (2nd ed.)* (p. 160):

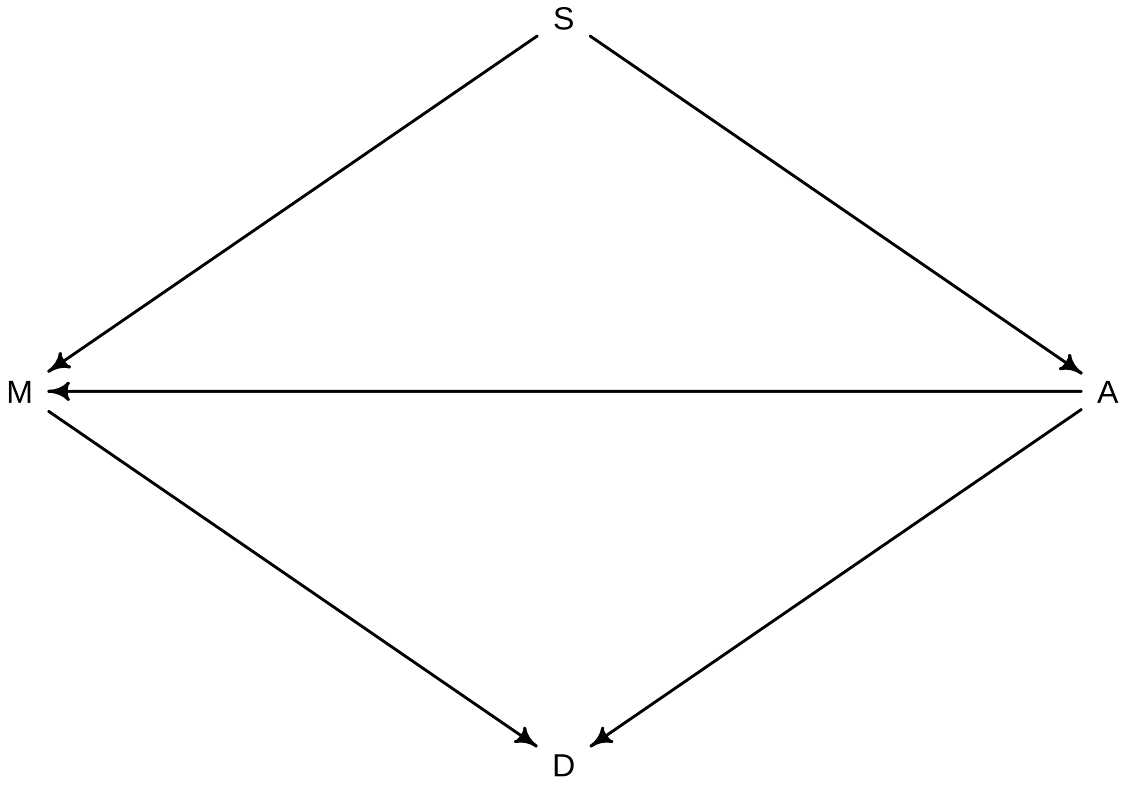
> **5H4.** Here is an open practice problem to engage your imagination. In the divorce data, states in the southern United States have many of the highest divorce rates. Add the `South` indicator variable to the analysis. First, draw one or more DAGs that represent your ideas for how Southern American culture might influence any of the other three variables (D, M, or A). Then list the testable implications of your DAGs, if there are any, and fit one or more models to evaluate the implications. What do you think the influence of "Southernness" is?

5H4.1 - Drawing the DAGS

```
# install.packages("dagitty")
library(dagitty)
```

```
## Warning: package 'dagitty' was built under R version 4.0.5
```

```
dag1 <- dagitty( "dag { M <- S -> A -> M -> D <- A}")
coordinates(dag1) <- list( x=c(D=0,S=0, M=-1, A=1), y=c(D=3, M=2, A=2, S=1))
drawdag( dag1 )
```



Considerations behind our choice of DAG:

S -> A: We believe that Southerness directly influences age of marriage due to several reasons. People from Southern states are known to be more religious and conservative. We believe that their prevalent positive view on the traditional marriage and their negative views on e.g. sex before marriage will cause people to marry younger. It is simply a social norm. Another practical issue is that there in many Southern states are more strict abortion rules, but as we all well know that does not stop people from having sex. So it is possible to imagine that if a woman gets pregnant, it is more likely that she will keep the child than a woman from a Nothern State, and this might force the young couple into marrying (again also because of social norms and concern for reputation).

S -> M: We believe that Southerness directly influences marriage rate because of the religious incentive to get married, as it is the "ideal" life according to Christianity.

A -> D: We believe that age influences divorce rate negatively, because of the assumption that the decision to marry was less well thought through by young people than older people, meaning that young people are more impulsive and driven by emotion, and therefor more relationships break. Also, young people simply have longer time to change and grow apart.

A -> M: We believe that age influences marriage rate negatively, because if people marry from a younger age there are simply more people to get married. So as the median age goes up, the marriage rate goes down.

M -> D: We believe that marriage rate influences divorce rate for the simple reason that more marriages increase the likelihood of more divorces.

These are the links that we thought were reasonable to assume, but you can of course make other DAGs. Some might argue that Southerness could have a direct impact on divorce rate, so it is not merely mediated by the links to age and marriage rate. Nevertheless, this is the DAG that we went with.

Testing Conditional Independencies:

```
impliedConditionalIndependencies(dag1)
```

```
## D _||_ S | A, M
```

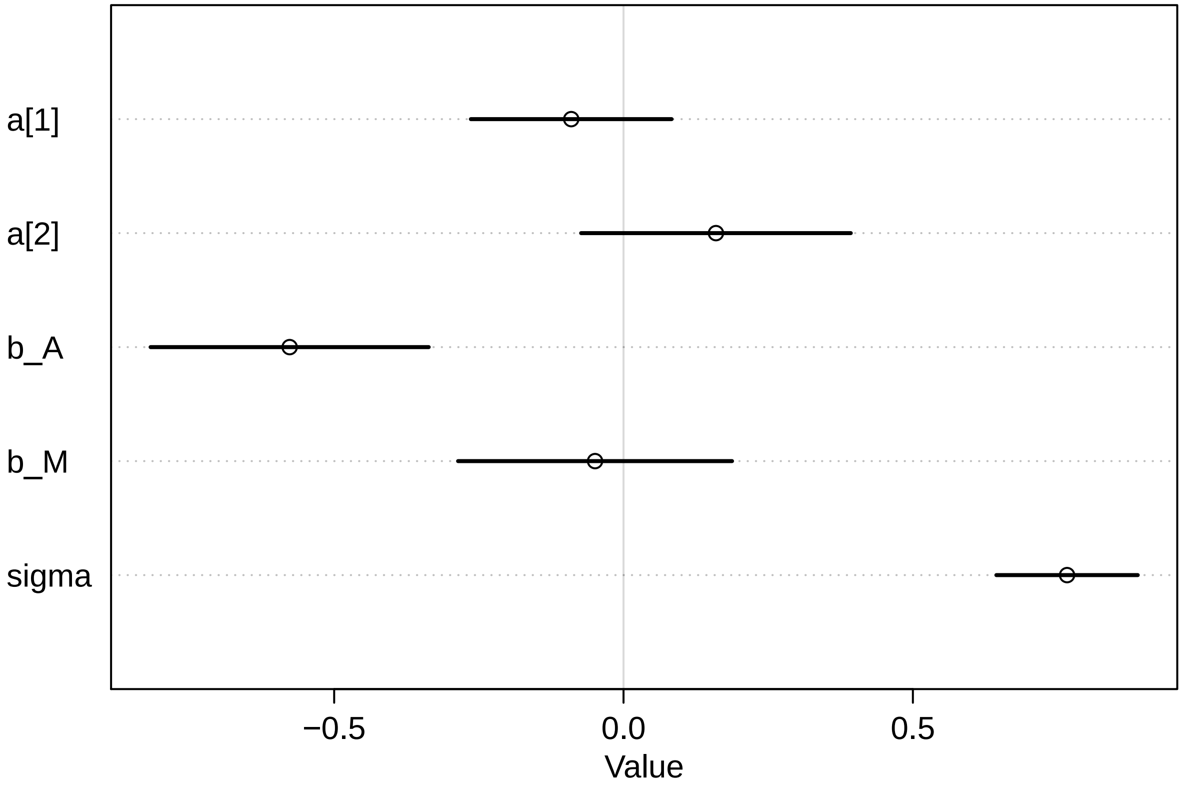
In terms of correlations in the model, we see that D is independent from S, conditional on A and M. Let's test this in our data:

Loading data:

```
data(WaffleDivorce)

d <- list()
d$A <- standardize( WaffleDivorce$MedianAgeMarriage )
d$D <- standardize( WaffleDivorce$Divorce )
d$M <- standardize( WaffleDivorce$Marriage )
d$S <- as.integer(WaffleDivorce$South+1)
```

Testing Conditional Independencies:

```
mD_SAM <- quap(
  alist(
   D ~ dnorm(mu, sigma),
   mu <- a[S] + b_A*A + b_M*M,
   a[S] ~ dnorm(0, 0.2),
   b_A ~ dnorm(0, 0.5),
   b_M ~ dnorm(0, 0.5),
   sigma ~ dexp(1)), data = d)

precis_plot(precis(mD_SAM, depth=2))
```



We wanted to see that S and D are independent (i.e. the effect of S on D is overlapping 0) when we stratify by A and M (i.e. include them in our model). This is exactly what we can see from the plot above.

5H4.2 - Making models to test implications of our causal reasoning:

Making model:

```r
# this model runs two regressions
m_D_A_M_S <- quap(
  alist(
    ## S -> A
    A ~ dnorm(mu_A, sigma_A),
    mu_A <- aA[S],
    aA[S] ~ dnorm(0,0.2),
    sigma_A ~ dexp(1),

    ## A -> D <- M
    D ~ dnorm( mu , sigma ) ,
    mu <- a + bM[S]*M + bA[S]*A ,
    a ~ dnorm( 0 , 0.2 ) ,
    bM[S] ~ dnorm( 0 , 0.5 ) ,
    bA[S] ~ dnorm( 0 , 0.5 ) ,
    sigma ~ dexp( 1 ),

    ## A -> M <- S
    M ~ dnorm( mu_M , sigma_M ),
    mu_M <- aM[S] + bAM[S]*A,
    aM[S] ~ dnorm( 0 , 0.2 ),
    bAM[S] ~ dnorm( 0 , 0.5 ),
    sigma_M ~ dexp( 1 )
    ) , data = d )

precis(m_D_A_M_S, depth=2)
```

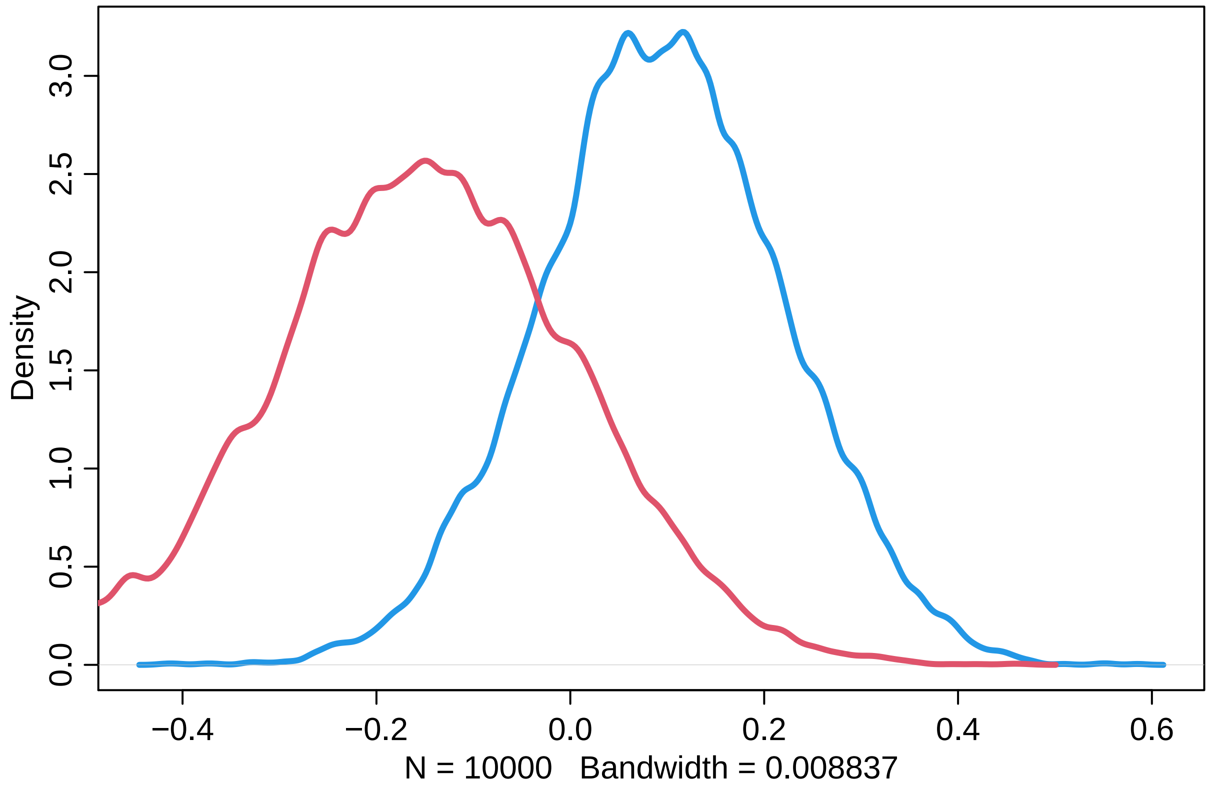
```
##                 mean         sd        5.5%      94.5%
## aA[1]     0.09338565 0.12514004 -0.10661231  0.2933836
## aA[2]    -0.14885848 0.15880166 -0.40265420  0.1049372
## sigma_A   0.95988800 0.09543649  0.80736206  1.1124139
## a        -0.04127333 0.09571905 -0.19425086  0.1117042
## bM[1]    -0.10531849 0.14866137 -0.34290807  0.1322711
## bM[2]     0.42396924 0.31818011 -0.08454403  0.9324825
## bA[1]    -0.53877157 0.15047099 -0.77925327 -0.2982899
## bA[2]    -0.83838169 0.31903953 -1.34826847 -0.3284949
## sigma     0.73027697 0.07369326  0.61250091  0.8480530
## aM[1]     0.04320631 0.09936658 -0.11560067  0.2020133
## aM[2]    -0.04535941 0.14730020 -0.28077358  0.1900548
## bAM[1]   -0.70253800 0.10196694 -0.86550085 -0.5395751
## bAM[2]   -0.54009160 0.27516963 -0.97986580 -0.1003174
## sigma_M   0.68019448 0.06807878  0.57139144  0.7889975
```

```
# M and A are strongly negatively associated. If we interpret this causally, it indicates that manipula
```

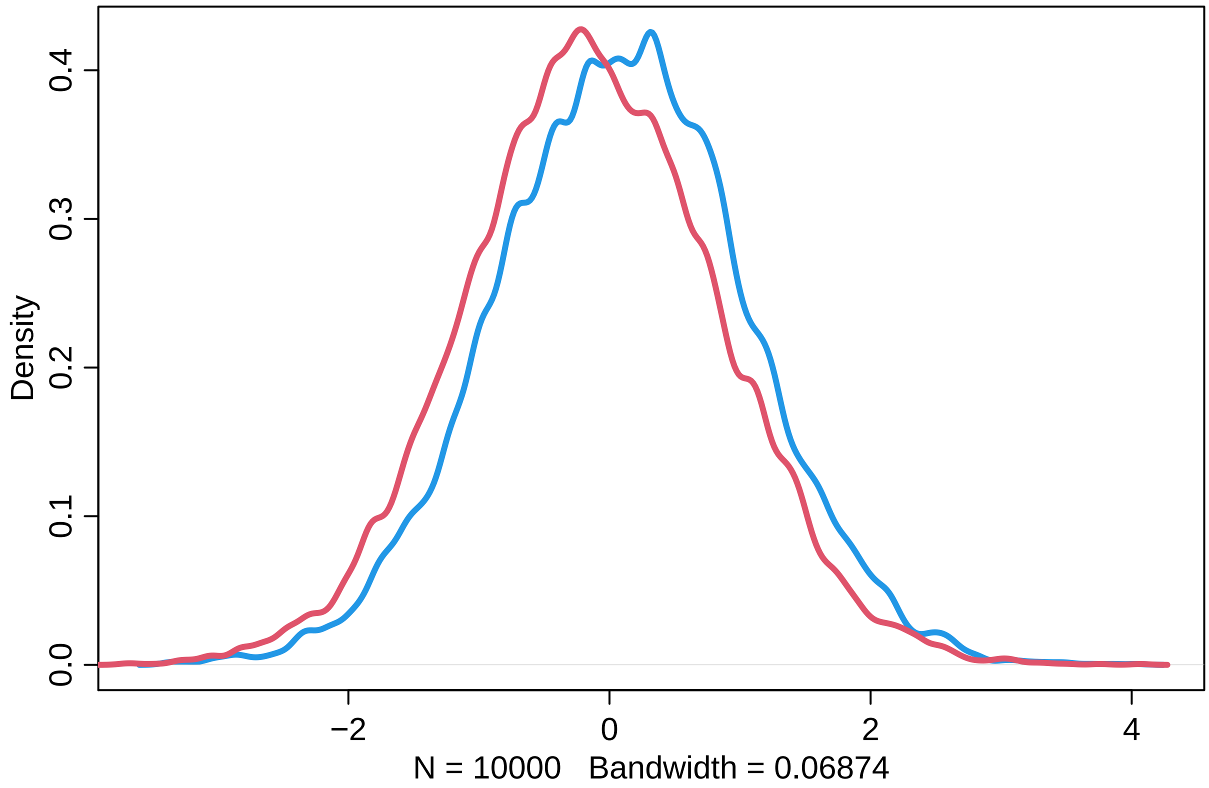
Now we have a bunch of parameters. In order to conclude on these, we are going to simulate from them and do contrasting. By looking at the parameters, we can get an idea about what could be interesting to dive into. As a rule of thumb, if you can multiply the standard deviation with two and add/subtract it from the mean, without the mean going to zero, then you might have an interesting effect. As we see, this is in most cases not possible with our model, but the two slopes for median age do have this trade. Therefor, we will start by investigating our claim that Southerness influences median marriage age.

1. investigation: How does Southerness influence median-marriage-age?

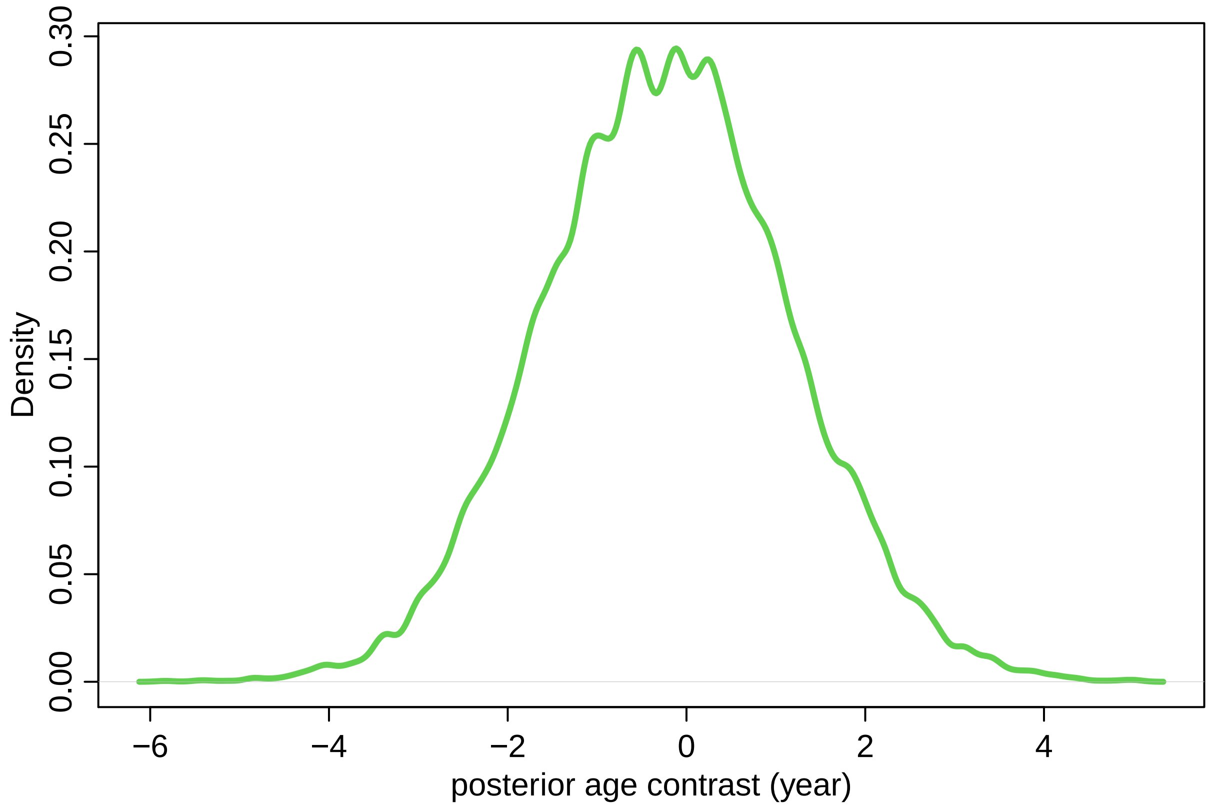
```
# simulating
post <- extract.samples(m_D_A_M_S)

dens(post$aA[,1], lwd=3, col=4) # blue = non-southern
dens(post$aA[,2], lwd=3, col=2, add=TRUE) # red = south
```



```
A1 <- rnorm(1e4, post$aA[,1], post$sigma_A) #simulating 10.000 observations and using the parameters fr
A2 <- rnorm(1e4, post$aA[,2], post$sigma_A) #same for Southerns
dens(A1, col=4, lwd=3) # non-south = blue
dens(A2, col=2, lwd=3, add=TRUE) # south = red
```



```
# contrast
A_contrast <- A2 - A1
dens(A_contrast,
     col=3, lwd=3,
     xlab="posterior age contrast (year)")
```



```
# proportion above zero
sum(A_contrast > 0)/1e4 # 42.64%
```

```
## [1] 0.4345
```

```
sum(A_contrast < 0)/1e4 # 57.36%
```

```
## [1] 0.5655
```

Answer: We see from the posterior predictive that there is a lot of overlap, but in order to be able to conclude anything, it is important to do the contrasting. From the contrast plot we see that the distribution is centered pretty close around zero, but it is skewed a little bit. More accurately, 57% of times you randomly pick a south-state, they have younger marriage age than non-south-state, based on our simulations. This means that Southerness is *not* a very good predictor of median-age, as this is very close to chance level.

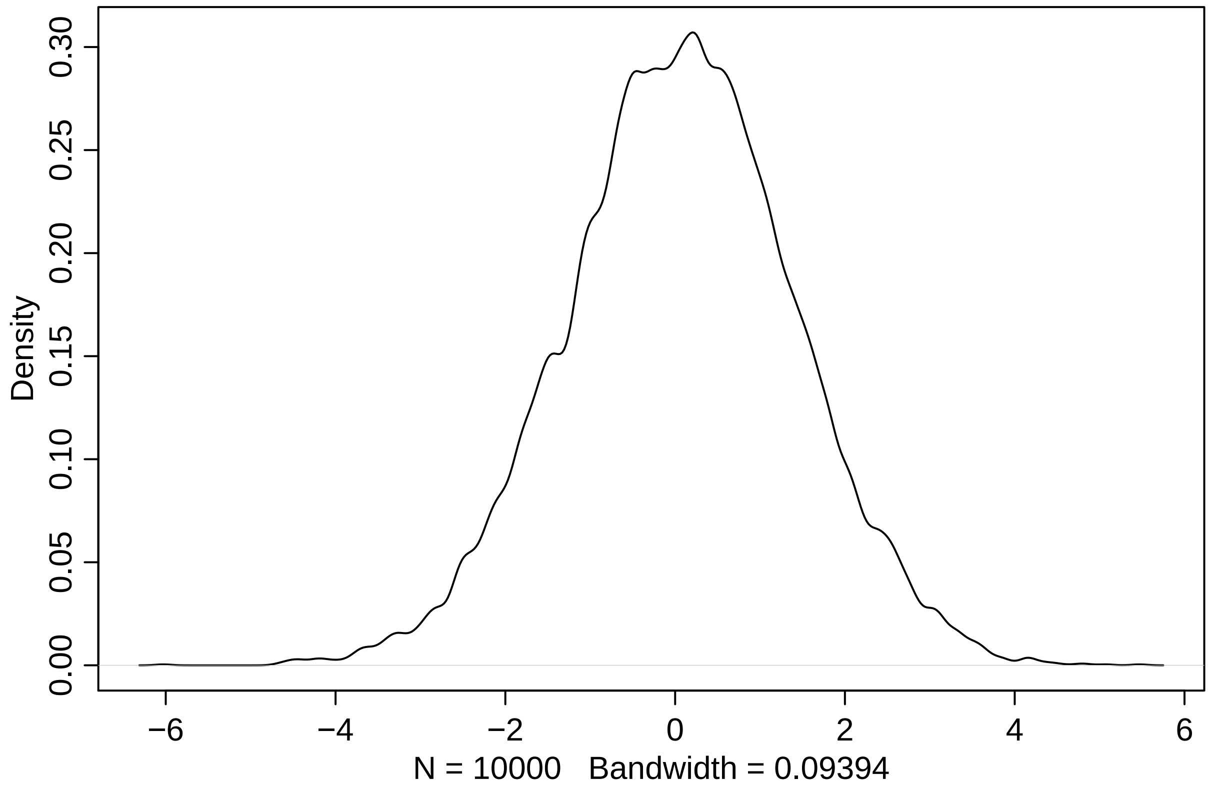
Next, we will investigate the other direct link from Southerness that we assumed in our DAG.

2. investigation: How does Southerners influence marriage rate?

```
m_D_A_M_S_sim <- sim(m_D_A_M_S,
             n = 1e4,
             data = list(S=c(1,2)),
             vars=c("A","M", "D"))
```

```
## Warning in if (left == var) {: the condition has length > 1 and only the first
## element will be used

## Warning in if (left == var) {: the condition has length > 1 and only the first
## element will be used

## Warning in if (left == var) {: the condition has length > 1 and only the first
## element will be used

## Warning in if (left == var) {: the condition has length > 1 and only the first
## element will be used
```

```
M_contrast <- m_D_A_M_S_sim$M[,2]-m_D_A_M_S_sim$M[,1]
dens(M_contrast)
```



```
sum(M_contrast > 0)/1e4 # 51.91%
```

```
## [1] 0.5241
```

```
sum(M_contrast < 0)/1e4 # 48.09%
```

```
## [1] 0.4759
```

Answer: Based on our contrasting, we estimate that 51.91% of the times you randomly pick a South state, their marriage rate will be higher than that of a non-southern state. You literally cannot get much closer to chance level, so we can conclude that Southerness is an extremely poor predictor of marriage rate.

Since we in our DAG assume that there is no direct causal link between Southerness and divorce rate, we should have encapsulated all the effect from by investigating the effect from Southerness on age and marriage rate. Since both of our analyses showed that Southerness was a very poor predictor, we conclude that Southerness does not have a big impact on a state's divorce rate.