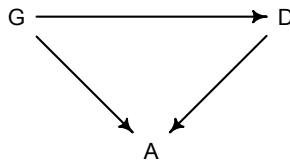


STATISTICAL RETHINKING WINTER 2020/2021
HOMEWORK, WEEK 6 SOLUTIONS

1. The implied DAG is:



where G is gender, D is discipline, and A is award. The direct causal effect of gender is the path $G \rightarrow A$. The total effect includes that path and the indirect path $G \rightarrow D \rightarrow A$. We can estimate the total causal influence (assuming this DAG is correct) with a model that conditions only on gender. I'll use a $N(-1,1)$ prior for the intercepts, because we know from domain knowledge that less than half of applicants get awards.

```
library(rethinking)
data(NWOWGrants)
d <- NWOWGrants

dat_list <- list(
  awards = as.integer(d$awards),
  apps = as.integer(d$applications),
  gid = ifelse( d$gender=="m" , 1L , 2L ) )

m1_total <- ulam(
  alist(
    awards ~ binomial( apps , p ),
    logit(p) <- a[gid],
    a[gid] ~ normal(-1,1)
  ), data=dat_list , chains=4 , cmdstan=TRUE )


precis(m1_total,2)
```

	mean	sd	5.5%	94.5%	n_eff	Rhat
a[1]	-1.53	0.06	-1.64	-1.43	1371	1
a[2]	-1.74	0.08	-1.88	-1.61	1291	1

Gender 1 here is male and 2 is female. So males have higher rates of award, on average. How big is the difference? Let's look at the contrast on absolute scale:

```
post <- extract.samples(m1_total)
diff <- inv_logit( post$a[,1] ) - inv_logit( post$a[,2] )
precis( list( diff=diff ) )
```

'data.frame': 2000 obs. of 1 variables:

	mean	sd	5.5%	94.5%	histogram
diff	0.03	0.01	0.01	0.05	

So a small 3% difference on average. Still, with such low funding rates (in some disciplines), 3% is a big advantage.

Now for the direct influence of gender, we condition on discipline as well:


```
dat_list$disc <- as.integer(d$discipline)
m1_direct <- ulam(
  alist(
    awards ~ binomial( apps , p ),
    logit(p) <- a[gid] + d[disc],
    a[gid] ~ normal(-1,1),
    d[disc] ~ normal(0,1) ),
  data=dat_list , chains=4 , cores=4 , cmdstan=TRUE )
precis(m1_direct,2)
```

	mean	sd	5.5%	94.5%	n_eff	Rhat4
a[1]	-1.33	0.30	-1.81	-0.84	207	1.01
a[2]	-1.47	0.31	-1.97	-0.97	208	1.02
d[1]	0.31	0.35	-0.25	0.88	280	1.01
d[2]	-0.01	0.33	-0.54	0.53	233	1.01
d[3]	-0.24	0.33	-0.75	0.29	242	1.01
d[4]	-0.28	0.35	-0.83	0.29	289	1.01
d[5]	-0.35	0.32	-0.85	0.16	231	1.01
d[6]	-0.04	0.35	-0.58	0.53	257	1.01
d[7]	0.27	0.38	-0.33	0.88	316	1.01
d[8]	-0.47	0.32	-0.95	0.03	223	1.01
d[9]	-0.22	0.33	-0.72	0.33	257	1.01

Those chains didn't sample very efficiently. This is because the model is over-parameterized—it has more parameters than absolutely necessary. This doesn't break it. It just makes the sampling less efficient. Anyway, now we can compute the gender difference again. On the relative scale:

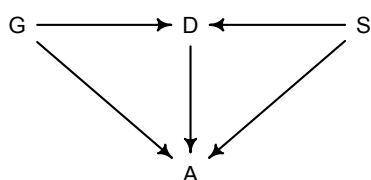
```
post <- extract.samples(m1_direct)
diff_a <- post$a[,1] - post$a[,2]
precis( list( diff_a=diff_a ) )
```

'data.frame': 6000 obs. of 1 variables:

	mean	sd	5.5%	94.5%	histogram
diff_a	0.14	0.11	-0.03	0.31	

Still an advantage for the males, but reduced and overlapping zero a bit. To see this difference on the absolute scale, we need to account for the base rates in each discipline as well. If you look at the `postcheck(m1_direct)` display, you'll see the predictive difference is very small. There are also several disciplines that reverse the advantage. If there is a direct influence of gender here, it is small, much smaller than before we accounted for discipline. Why? Because again the disciplines have different funding rates and women apply more to the disciplines with lower funding rates. But it would be hasty, I think, to conclude there are no other influences. There are after all lots of unmeasured confounds...

2. The implied DAG is:



where S is stage of career (unobserved). This DAG has the same structure as the grandparents-parents-children-neighborhoods example from earlier in the course. When we condition on discipline D it opens a backdoor path through S to A. It is not possible here to get an unconfounded estimate of gender on awards.

Here's a simulation to demonstrate the potential issue.

```

set.seed(1913)
N <- 1000
G <- rbern(N)
S <- rbern(N)
D <- rbern( N , p=inv_logit( G + S ) )
A <- rbern( N , p=inv_logit( 0.25*G + D + 2*S - 2 ) )
dat_sim <- list( G=G , D=D , A=A )

```

This code simulates 1000 applicants. There are 2 genders (G 0/1), 2 stages of career (S 0/1), and 2 disciplines (D 0/1). Discipline 1 is chosen more by gender 1 and career stage 1. So that could mean more by males and later stage of career. Then awards A have a consistent bias towards gender 1, and discipline 1 has a higher award rate, and stage 1 also a higher award rate. If we analyze these data:

```

m2_sim <- ulam(
  alist(
    A ~ bernoulli(p),
    logit(p) <- a + d*D + g*G,
    c(a,d,g) ~ normal(0,1)
  ), data=dat_sim , chains=4 , cores=4 , cmdstan=TRUE )
precis(m2_sim)

```

```

mean    sd   5.5% 94.5% n_eff Rhat4

```

```
g  0.34 0.13  0.14  0.56 1127    1
d  1.22 0.15  0.98  1.47  743    1
a -1.08 0.14 -1.30 -0.85  742    1
```

The parameter g is the advantage of gender 1. It is much smaller than the true advantage, even with 1000 applicants.

It is also possible to have no gender influence and infer it by accident. Try these settings:

```
set.seed(1913)
N <- 1000
G <- rbern(N)
S <- rbern(N)
D <- rbern( N , p=inv_logit( 2*G - S ) )
A <- rbern( N , p=inv_logit( 0*G + D + S - 2 ) )
dat_sim2 <- list( G=G , D=D , A=A )
m2_sim_2 <- ulam( m2_sim , data=dat_sim2 , chains=4 , cores=4 , cmdstan=TRUE )
precis(m2_sim_2,2)
```

```
      mean    sd  5.5% 94.5% n_eff Rhat4
g  0.25 0.16  0.00  0.51  1081  1.00
d  0.27 0.16  0.02  0.52   948  1.01
a -1.13 0.13 -1.34 -0.92   908  1.01
```

Now it looks like gender 1 has a consistent advantage, but in fact there is no advantage in the simulation.

What can be done in these situations? The best thing is to stop analyzing convenient, confounded data and instead design a real study.

3. First, we need to make some 0/1 dummy variables for the categorical variables P , A , and V in the data frame. Name them whatever you like, but here are my choices:

```
library(rethinking)
library(MASS)
data(eagles)
d <- eagles
d$pirateL <- ifelse( d$P=="L" , 1 , 0 )
d$victimL <- ifelse( d$V=="L" , 1 , 0 )
d$pirateA <- ifelse( d$A=="A" , 1 , 0 )
```

Now to fit a model. I'm going to use the standard intercept prior here, and a slightly narrower prior for each coefficient. This allows for large effects, but does regularize. For the predictor structure, let's allow an interaction between the size of the pirate and the size of the victim. Let's also include an interaction of victim size with pirate age.

```
f <- alist(
  y ~ dbinom( n , p ),
```

```

logit(p) <- a + bP*pirateL + bV*victimL + bA*pirateA +
            bPV*pirateL*victimL + bAV*pirateA*victimL ,
a ~ dnorm(0,1.5),
c(bP,bV,bA,bPV,bAV) ~ dnorm(0,1) )

m1 <- ulam( f , data=d , chains=4 , cmdstan=TRUE )

precis(m1)

```

	mean	sd	5.5%	94.5%	n_eff	Rhat4
a	0.17	0.52	-0.66	0.99	730	1.00
bAV	-0.91	0.63	-1.89	0.11	871	1.00
bPV	0.42	0.66	-0.63	1.45	759	1.00
bA	1.44	0.55	0.57	2.35	802	1.00
bV	-2.65	0.61	-3.65	-1.66	744	1.01
bP	2.51	0.59	1.54	3.43	917	1.00

Now for interpreting these estimates. It's really not easy to interpret the coefficients directly. This is typical for interaction models. But you can probably see that the interaction coefficient bPV has a good amount of probability on both sides of zero. It's not clear that there is any important interaction. The bAV interaction is stronger. The other thing you can see is that the coefficients bV and bP are large. These variables do matter.

Let's just plot. There are many ways to plot the posterior predictions. I'll use a straightforward format with cases on the horizontal and probability/count on the vertical. Here's the code for the first plot, showing proportion success on the vertical axis:

```

d$psuccess <- d$y / d$n

p <- link(m1)
y <- sim(m1)

p.mean <- apply( p , 2 , mean )
p.PI <- apply( p , 2 , PI )
y.mean <- apply( y , 2 , mean )
y.PI <- apply( y , 2 , PI )

# plot raw proportions success for each case
plot( d$psuccess , col=rainbow(2) ,
      ylab="successful proportion" , xlab="case" , xaxt="n" ,
      xlim=c(0.75,8.25) , pch=16 )

# label cases on horizontal axis
axis( 1 , at=1:8 ,
      labels=c( "LAL","LAS","LIL","LIS","SAL","SAS","SIL","SIS" ) )

```

```
# display posterior predicted proportions successful
points( 1:8 , p.mean )
for ( i in 1:8 ) lines( c(i,i) , p.PI[,i] )
```

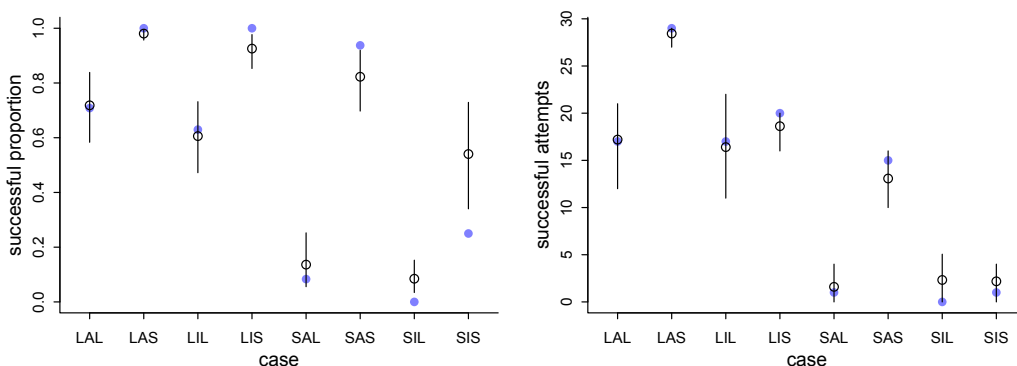
And here's the second plot, showing number of successes on the vertical:

```
# plot raw counts success for each case
plot( d$y , col=range(2 ,
  ylab="successful attempts" , xlab="case" , xaxt="n" ,
  xlim=c(0.75,8.25) , pch=16 )

# label cases on horizontal axis
axis( 1 , at=1:8 ,
  labels=c( "LAL","LAS","LIL","LIS","SAL","SAS","SIL","SIS" ) )

# display posterior predicted successes
points( 1:8 , y.mean )
for ( i in 1:8 ) lines( c(i,i) , y.PI[,i] )
```

And here are the resulting plots:



So what's different? The biggest difference is probably that the left plot (proportions) makes the probabilities more comparable, because it ignores the sample size for each case on the horizontal axis. This has the advantage of showing, for example, that SIS attempts are predicted to be somewhat successful, even though only 4 of them were observed. The right plot (counts), in contrast, makes it hard to see the differing probabilities, because samples size varies so much across cases.

On the other hand, the count plot (right) has the advantage of showing additional uncertainty that arises from the binomial process. Inspect the SIS case again, for example. The observed proportion of SIS successes is outside and below the probability interval (left plot). However, in the right plot, the model can accommodate the SIS cases, due to additional uncertainty arising from the binomial process. In

other words, an additional level of stochasticity is factored into the right plot, and so in total looking also at counts might be necessary to seriously critique the model.

Now what about the interactions? The coefficient b_{AV} is negative. Does this mean older pirates are disadvantaged against large victims? No. Look at the plots. Adult pirates (A) are better than immature (I), holding other variables the same. The coefficient is a gear in a complex machine. Don't interpret it. Plot and interpret on the outcome space.