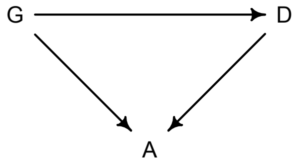
STATISTICAL RETHINKING WINTER 2019
HOMEWORK, WEEK 6 SOLUTIONS

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.

```
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 )
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 (penguin) 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 , iter=3000 )
precis(m1_direct,2)
```

```
      mean   sd  5.5% 94.5% n_eff Rhat
a[1] -1.33 0.31 -1.84 -0.85   615 1.01
a[2] -1.47 0.31 -1.98 -0.98   636 1.01
d[1]  0.31 0.36 -0.26  0.90   848 1.01
d[2] -0.01 0.33 -0.54  0.54   722 1.01
d[3] -0.24 0.33 -0.75  0.30   694 1.01
d[4] -0.28 0.36 -0.85  0.31   791 1.00
d[5] -0.35 0.33 -0.86  0.19   691 1.01
d[6] -0.03 0.35 -0.58  0.53   789 1.00
d[7]  0.28 0.39 -0.33  0.91   968 1.00
d[8] -0.46 0.32 -0.96  0.07   669 1.01
d[9] -0.21 0.34 -0.74  0.34   758 1.01
```

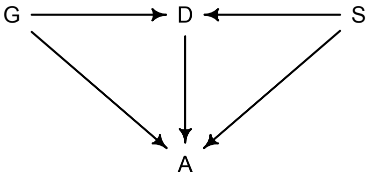
Those chains didn't sample very efficiently. This likely because the model is over-parameterized—it has more parameters than absolutely necessary. This doesn't break it. It just makes the sampling less efficient. Anyway, now we can compute the gender difference again. On the relative scale:

```
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 , chains=4 , cores=4 )
precis(m2_sim)
```

```
   mean   sd  5.5% 94.5% n_eff Rhat
g  0.09 0.13 -0.13  0.30   984    1
d  1.21 0.15  0.98  1.46   927    1
a -0.90 0.13 -1.12 -0.70   983    1
```

The parameter g is the advantage of gender 1. It is smaller than the true advantage and the estimate straddles zero quite a lot, 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 )
precis(m2_sim_2,2)
```

```
   mean   sd  5.5% 94.5% n_eff Rhat
g  0.25 0.15  0.00  0.48  1153    1
d  0.28 0.15  0.03  0.52  1036    1
a -1.12 0.12 -1.31 -0.94  1166    1
```

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

3. First let's load the data and set it up for use:

```
library(rethinking)
data(Primates301)
d <- Primates301
d2 <- d[ complete.cases( d$social_learning , d$brain , d$research_effort ) , ]
dat <- list(
    soc_learn = d2$social_learning,
    log_brain = standardize( log(d2$brain) ),
    log_effort = log(d2$research_effort)
)
```

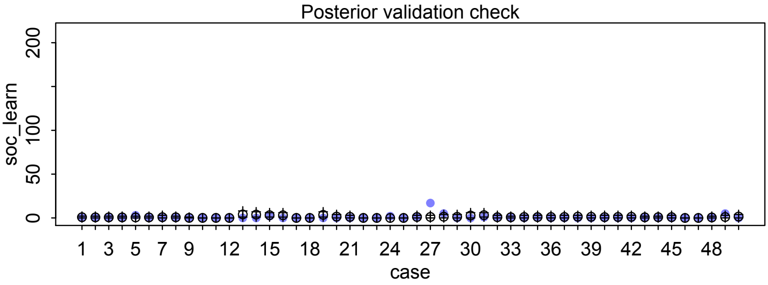
Now we first want a model with social learning as the outcome and brain size as a predictor. For this Poisson GLM, I'm going to use a N(0,1) prior on the intercept, since we know the counts should be small.

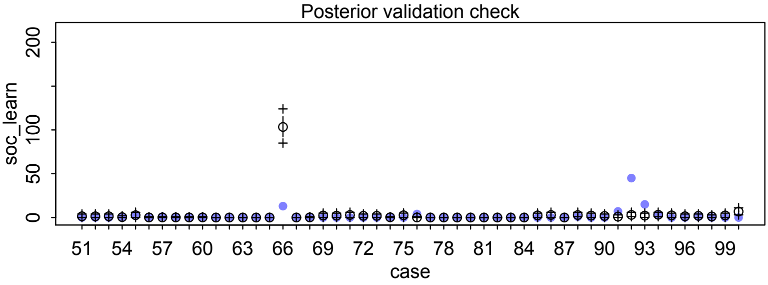
```
m3_1 <- ulam(
    alist(
        soc_learn ~ poisson( lambda ),
        log(lambda) <- a + bb*log_brain,
        a ~ normal(0,1),
        bb ~ normal(0,0.5)
    ), data=dat , chains=4 , cores=4 )
precis( m3_1 )
```

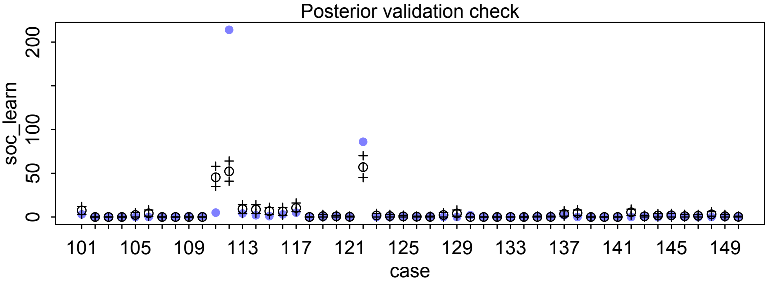
```
    mean   sd  5.5% 94.5% n_eff Rhat
```

```
a   -1.18 0.12 -1.36 -0.99   423    1
bb   2.76 0.08  2.64  2.88   445    1
```

Brain size seems to be strongly associated with social learning observations. Let's look at the posterior predictions. I'll use `postcheck(m3_1,window=50)`:







The blue points are the raw data, recall. These are not great posterior predictions. Clearly other factors are in play.

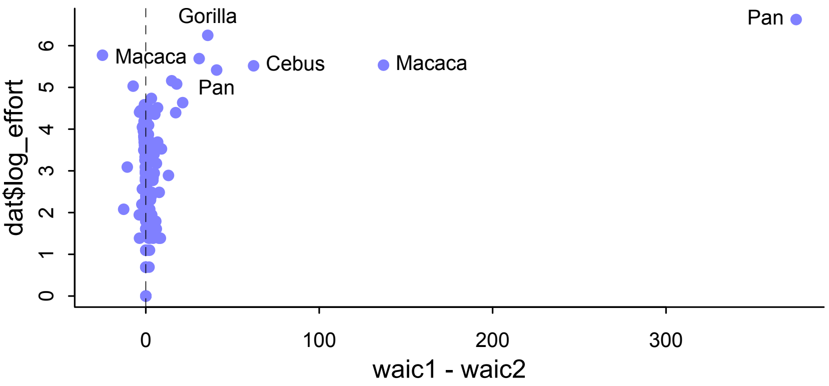
Let's try the research effort variable now:

```
m3_2 <- ulam(
    alist(
        soc_learn ~ poisson( lambda ),
        log(lambda) <- a + be*log_effort + bb*log_brain,
        a ~ normal(0,1),
        c(bb,be) ~ normal(0,0.5)
    ), data=dat , chains=4 , cores=4 )
precis( m3_2 )
```

```
    mean   sd  5.5% 94.5% n_eff Rhat
a  -5.97 0.33 -6.49 -5.45   479 1.01
be  1.53 0.07  1.42  1.64   456 1.01
bb  0.46 0.08  0.33  0.59   599 1.01
```

Brain size bb is still positively associated, but much less. Research effort be is strongly associated. To see how these models disagree, let's use pointwise WAIC to see which cases each predicts well.

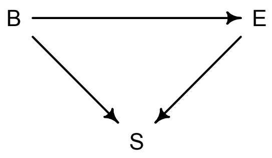
```
waic1 <- WAIC( m3_1 , pointwise=TRUE )
waic2 <- WAIC( m3_2 , pointwise=TRUE )
plot( waic1 - waic2 , dat$log_effort , col=rangi2 , pch=16 )
identify( waic1-waic2 , dat$log_effort , d2$genus , cex=0.8 )
abline(v=0,lty=2,lwd=0.5)
```



Species on the right of the vertical line fit better for model m3_2, the model with research effort. These are mostly species that are studied a lot, like chimpanzees (*Pan*) and macaques (*Macaca*). The genus *Pan* especially has been a focus on social learning research, and its counts are inflated by this.

This is a good example of how the nature of measurement influences inference. There are likely a lot of false zeros in these data, species that are not studied often enough to get a good idea of their learning tendencies. Meanwhile every time a chimpanzee sneezes, someone writes a social learning paper.

Okay, finally I asked for a DAG. This is my guess:



B is brain size, E is research effort, and S is social learning. Research effort doesn't actually influence social learning, but it does influence the value of the variable. The model results above are consistent with this DAG in the sense that including E reduced the association with B, which we would expect when we close the indirect path through E. If researchers choose to look for social learning in species with large brains, this leads to an exaggerated estimate of the association between brains and social learning.