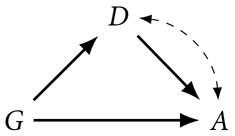
STATISTICAL RETHINKING 2025
WEEK 5 SOLUTIONS

1. The implied DAG is:



where G is gender, D is discipline, and A is award. The dashed curve between D and A represents possible confounds. The direct causal effect of gender is the path $G \rightarrow A$. The total effect of G 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. This prior is certainly too wide, but it is *weakly* informative instead of strongly information. This means it contains the most important information we have about the parameters, but it is weak enough to allow the data to surprise us and reveal potential model misspecifications.

```
library(rethinking)
data(NWOGrants)
d <- NWOGrants
dat <- list(
    A = as.integer(d$awards),
    N = as.integer(d$applications),
    G = ifelse( d$gender=="f" , 1L , 2L ) ,
    D = as.integer(d$discipline)
)

# for total effect, just G, no D
m1 <- ulam(
    alist(
        A ~ binomial( N , p ),
        logit(p) <- a[G],
        a[G] ~ normal(-1,1)
    ), data=dat , chains=4 , cores=4 )

precis(m1,2)
```

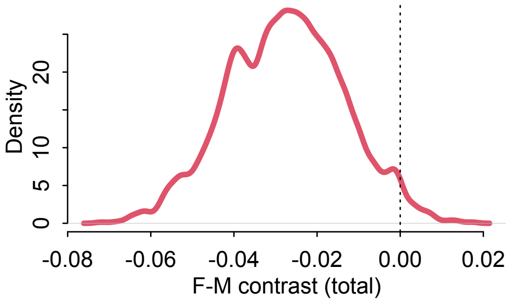
```
      mean   sd  5.5% 94.5% rhat ess_bulk
a[1] -1.74 0.08 -1.87 -1.61    1  1213.46
```

```
a[2] -1.53 0.06 -1.64 -1.43    1  1484.56
```

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

```
post1 <- extract.samples(m1)
post1$pG1 <- inv_logit( post1$a[,1] )
post1$pG2 <- inv_logit( post1$a[,2] )
post1$G_contrast <- post1$pG1 - post1$pG2

dens( post1$G_contrast , lwd=4 , col=2 , xlab="F-M contrast (total)" )
abline( v=0 , lty=3 )
```



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

What sort of intervention does this estimate reference? You could imagine an intervention on applicant gender directly. That is what the calculation seems to suggest. However that kind of intervention is implausible. It could instead be the perception of gender by the people reading the applications. There is no right answer here. The point is just to think carefully about the meaning of the causal estimate in terms of what kind of counter-factual (not actual) intervention it might refer to. Even if the intervention is impossible (changing everyone's gender e.g.), the causal estimate could be correct. But then we have the additional problem of thinking about mechanisms and which other interventions are possible.

Another variable that people routinely produce causal estimates for is age. It is hard to imagine an intervention on age. So what can such a causal estimate mean? I won't give an answer. But it's worth thinking about it.

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

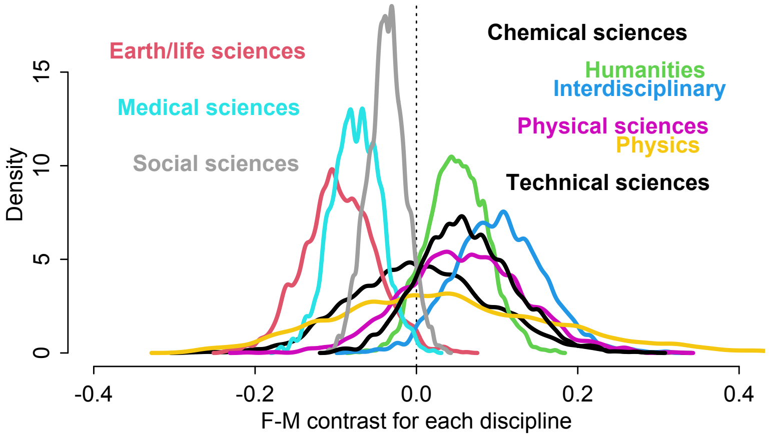
```
m2 <- ulam(
    alist(
        A ~ binomial( N , p ),
        logit(p) <- a[G,D],
        matrix[G,D]:a ~ normal(-1,1)
    ), data=dat , chains=4 , cores=4 )
```

```
precis(m2,3)
```

```
        mean   sd  5.5% 94.5% rhat ess_bulk
a[1,1] -1.08 0.34 -1.65 -0.56 1.00  2896.21
a[2,1] -1.02 0.24 -1.42 -0.65 1.01  3160.91
a[1,2] -1.76 0.26 -2.16 -1.38 1.01  3090.12
a[2,2] -1.13 0.18 -1.41 -0.85 1.01  2766.87
a[1,3] -1.43 0.19 -1.75 -1.13 1.00  3394.58
a[2,3] -1.77 0.19 -2.08 -1.48 1.00  2870.17
a[1,4] -1.28 0.27 -1.72 -0.85 1.00  2814.45
a[2,4] -1.99 0.28 -2.45 -1.56 1.00  3497.24
a[1,5] -2.05 0.19 -2.36 -1.74 1.00  3108.61
a[2,5] -1.46 0.16 -1.72 -1.20 1.01  3543.41
a[1,6] -1.21 0.36 -1.79 -0.66 1.00  3349.71
a[2,6] -1.42 0.21 -1.76 -1.10 1.00  2928.57
a[1,7] -1.21 0.63 -2.24 -0.25 1.00  3043.93
a[2,7] -1.02 0.27 -1.45 -0.61 1.00  2460.68
a[1,8] -2.02 0.16 -2.28 -1.78 1.00  2588.39
a[2,8] -1.70 0.13 -1.91 -1.50 1.00  2895.53
a[1,9] -1.32 0.30 -1.81 -0.86 1.00  2403.49
a[2,9] -1.65 0.19 -1.96 -1.35 1.00  2992.78
```

It isn't possible to make any sense out of this table. But we can compute the contrast in each discipline to see what is going on:

```
# show contrasts for each discipline
plot( NULL , xlim=c(-0.4,0.4) , ylim=c(0,18) ,
    xlab="F-M contrast for each discipline" , ylab="Density" )
abline( v=0 , lty=3 )
dat$disc <- as.character(d$discipline)
disc <- dat$disc[order(dat$D)]
for ( i in 1:9 ) {
    pG1Di <- link(m2,data=list(D=i,N=1,G=1))
    pG2Di <- link(m2,data=list(D=i,N=1,G=2))
    Gcont <- pG1Di - pG2Di
    dens( Gcont , add=TRUE , lwd=3 , col=i )
    xloc <- ifelse( mean(Gcont) < 0 , -0.35 , 0.35 )
    xpos <- ifelse( mean(Gcont) < 0 , 4 , 2 )
    text( xloc + 0.5*mean(Gcont) , 18-i , disc[2*i] , col=i ,
        pos=xpos , font=2 )
}
```



Not the nicest looking figure ever. But it shows the variation across disciplines, with some showing higher rates for women (right) and others for men (left).

Keep in mind when interpreting these estimates that the direct effect of gender is potentially (probably?) confounded by hidden differences in applicant ability or preparation. The dashed confound edge in the DAG is there to remind us that we can't exclude the possibility that who choses to apply within each discipline is associated with other traits that influence quality of applications. For example, women who apply in a field in which they anticipate discrimination may work harder on their applications, which in turn could hide or even reverse the effect of discrimination. Other kinds of data, such as details of the writing or review process.

Okay, now to compute the average direct effect of gender, imagining an intervention that changes perception of G on each grant application but does not alter D. First let's count up all of the applications and count them also by each discipline.

```
total_apps <- sum(dat$N)
apps_per_disc <- sapply( 1:9 , function(i) sum(dat$N[dat$D==i]) )
```

Next we simulate the same number of applications, to the same disciplines, but changing G to 1 in each case:

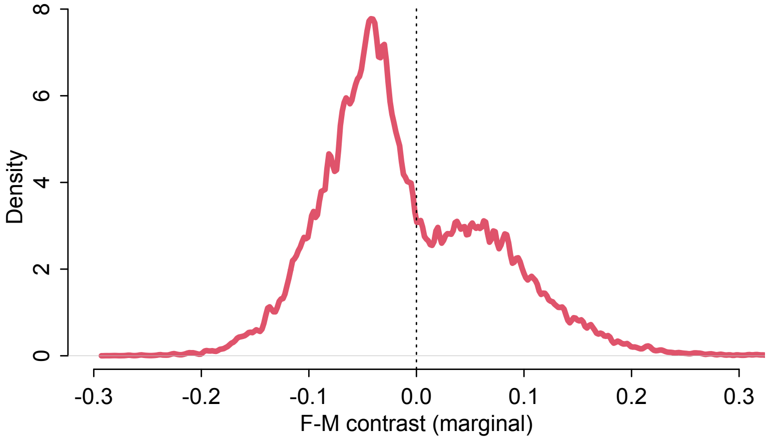
```
pG1 <- link(m2,data=list(
    D=rep(1:9,times=apps_per_disc),
    N=rep(1,total_apps),
    G=rep(1,total_apps)))
```

And the same but with $G = 2$ for all applications:

```
pG2 <- link(m2,data=list(
    D=rep(1:9,times=apps_per_disc),
    N=rep(1,total_apps),
    G=rep(2,total_apps)))
```

We plot the contrast distribution:

```
dens( pG1 - pG2 , lwd=4 , col=2 , xlab="F-M contrast (marginal)" ,
    xlim=c(-0.3,0.3) )
abline( v=0 , lty=3 )
```



The expected direct effect varies by discipline, and so when we post-stratify by discipline, the distribution of expected effects is broad. It is important to note that this density is not the same as the density of the total causal effect of G, but the total effect includes any causal effect on G on D. When we remove that (statistically), the result is that the advantages are rather balanced, with the mean of the density above being almost zero and the 89% interval about -0.12 to 0.12.