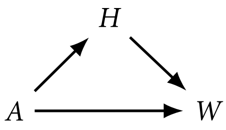
STATISTICAL RETHINKING 2024
WEEK 2 SOLUTIONS

1. The DAG you need is:



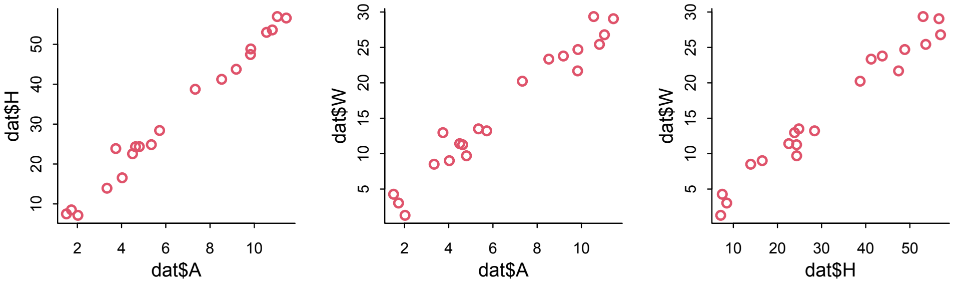
To turn this DAG into a generative model, we simulate *H* and *W* from values of *A*. We also need to make assumptions about how *A* influences *H* and *W*, as well as how *H* influences *W*. Here is a template function, based on the example from the book/lecture.

```
sim_HW <- function(A,bAH=5,bHW=0.5,bAW=0.1) {
    N <- length(A) # number of individuals
    H <- rnorm(N,bAH*A,2)
    W <- rnorm(N,bHW*H+bAW*A,2)
    data.frame(A,H,W)
}
```

The important thing is not the values I've chosen for the causal effects. The important thing is to get the order of simulation right. And that is done by always simulating the variables without any parents (causes) first. Then those variables with causes you have already simulated, and so on, until you've simulated all of the variables. In this case, we need to simulate *H* before we can simulate *W*, since *W* depends upon *H*.

Let's make a example synthetic sample and plot it, to see how these simulated people look. Remember, we are considering only ages under 13.

```
dat <- sim_HW( runif(20,1,12) )
plot( dat$A , dat$H , lwd=2 , col=2 )
plot( dat$A , dat$W , lwd=2 , col=2 )
plot( dat$H , dat$W , lwd=2 , col=2 )
```

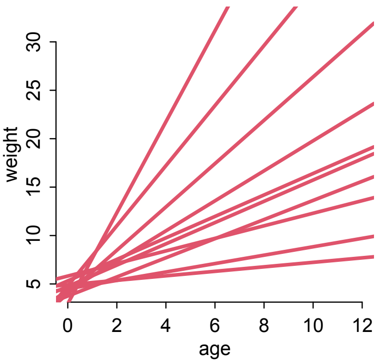


These relationships are not very realistic. The simulated children are too short, and young kids also do not grow in a linear fashion. But the simulation is structured like the DAG at least.

2. Since we want the total effect of age, we just need a linear regression of weight on age. Let's set up the data and then simulate some priors.

```
library(rethinking)
data(Howell1)
d <- Howell1
d <- d[ d$age < 13 , ]

# sim from priors
n <- 10
a <- rnorm(n,5,1)
b <- runif(n,0,10)
plot( NULL , xlim=range(d$age) , ylim=range(d$weight) ,
    xlab="age" , ylab="weight" )
for ( i in 1:n ) abline( a[i] , b[i] , lwd=3 , col=2 )
```



These were my first guess, given that the relationship must be positive and that weight at age zero is birth weight, and average birth weight is around 5 kilograms (but varies a lot).

Here's the model.

```
m2 <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + b*A,
        a ~ dnorm(5,1),
        b ~ dunif(0,10),
        sigma ~ dexp(1)
    ), data=list(W=d$weight,A=d$age) )
precis(m2)
```

```
      mean   sd 5.5% 94.5%
a     7.17 0.34 6.62  7.71
b     1.38 0.05 1.29  1.46
sigma 2.51 0.15 2.28  2.74
```

The total causal effect of each year of growth is given (in this case) by the parameter `b`. So its 89% interval is 1.29 to 1.46 kilograms per year. There is nothing to marginalize in this case, because there are no covariates.

3 - OPTIONAL CHALLENGE. There are two tasks here. The first is to convert the data from height measurement to increments in height. The second is to model the increments so they are always positive.

To convert the data to increments, we just subtract each height from the previous height for the same child. This means that the first occasion of measurement has no increment. So we start with the second occasion. There are lots of ways to do this in code. Here is how I did it.

```
data(Oxboys)
d <- Oxboys

d$delta <- NA
for ( i in 1:nrow(d) ) {
    if ( d$Occasion[i] > 1 )
        d$delta[i] <- d$height[i] - d$height[i-1]
}
d <- d[ !is.na(d$delta) , ]
```

The data frame `d` now has a new column `delta` with the increments. And I deleted the first occasion for each boy, since we won't model it.

Now we need a statistical model. There are a few ways to constrain the distribution of the increments to be positive. The easy way is to think of them as log-normal measurements. So you could log them first and them do an ordinary linear regression with them. You just need to exponentiate them later to get them back on the right scale. Or you can use a log-normal regression. I'll do that. Here's the model code.

```
m4 <- quap(
    alist(
        delta ~ dlnorm( alpha , sigma ),
        alpha ~ dnorm( 0 , 0.1 ),
        sigma ~ dexp( 3 )
    ), data=list(delta=d$delta) )
```

I use a log-normal distribution for the delta values. The trick with log-normal distributions is that the parameters refer to the log distribution. So alpha above is the mean of a normal distribution, not a log-normal distribution. Confusing, I know. The mean of the log-normal we are estimating is $\exp(\alpha + \sigma^2/2)$.

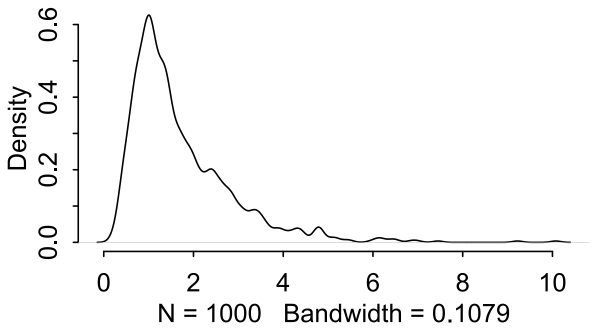
If you play around with prior predictive simulation, you'll see that variance in the priors leads to really explosive means. Here is the code I used:

```
# simulation from priors
n <- 1e3
alpha <- rnorm(n,0,0.1)
sigma <- rexp(n,3)
delta_sim <- rlnorm(n,alpha,sigma)
dens(delta_sim)
```

If you let σ be wider, it will make the prior mean way too high. This is typical of the log-normal. Normal distributions are nice and well-behaved. Log-normal distributions are not. They are tinder boxes.

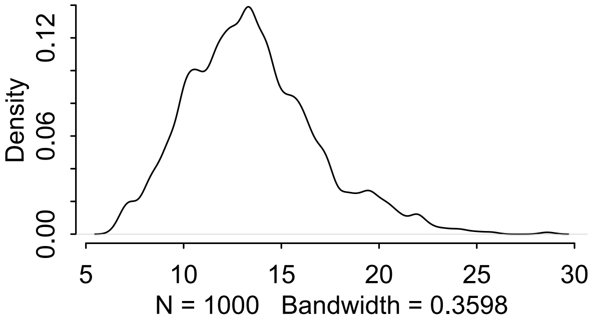
Okay, we have the posterior from `m4`. Now we can plot the posterior distribution of increments and their sum over 8 occasions. First the increment distribution:

```
post <- extract.samples(m4)
dsim <- rlnorm(1e3,post$alpha,post$sigma)
dens(dsim)
```



And the sum over 8 occasions of growth is:

```
inc_sum <- sapply( 1:1000 ,
    function(s) sum(rlnorm(8,post$alpha[s],post$sigma[s])) )
dens(inc_sum)
```



A source of variation that we have ignored is variation among boys in their growth rates. Later in the course, you'll see how to deal with this.