Exercises week 3

prof. Richard Torkar

March 28, 2021

> Information is the resolution of uncertainty.
>
> Claude Shannon

The weekly exercises are mainly from McElreath's notes which can be found on GitHub: `https://github.com/rmcelreath/stat_rethinking_2020/tree/main/homework`

Exercise 1. Consider three fictional Polynesian islands. On each there is a Royal Ornithologist charged by the king with surveying the birb population. They have each found the following proportions of 5 important birb species:

	Birb *A*	Birb *B*	Birb *C*	Birb *D*	Birb *E*
Island 1	0.2	0.2	0.2	0.2	0.2
Island 2	0.8	0.1	0.05	0.025	0.025
Island 4	0.05	0.15	0.7	0.05	0.05

Notice that each row sums to 1, all the birbs. This problem has two parts. It is not computationally complicated. But it is conceptually tricky. First, compute the entropy of each island's birb distribution. Interpret these entropy values.

Second, use each island's birb distribution to predict the other two. This means to compute the K-L Divergence of each island from the others, treating each island as if it were a statistical model of the other islands. You should end up with 6 different K-L Divergence values. Which island predicts the others best? Why?

Exercise 2. Recall the marriage, age, and happiness collider bias example from Chapter 6. Run models `m6.9` and `m6.10` again. Compare these two models using WAIC (or LOO, they will produce identical results). Which model is expected to make better predictions? Which model provides the correct causal inference about the influence of age on happiness? Can you explain why the answers to these two questions disagree?

SOLUTION 1. To compute the entropies, we just need a function to compute the entropy. Information entropy, as defined in lecture and the book, is simply:

$$H(p) = -\sum_i p_i log(p_i)$$

where p is a vector of probabilities summing to 1. In R code this would look like:

```
H <- function(p) -sum(p*log(p))
```

I'll make a list of the birb distributions and then push each through the function above.

```
IB <- list()
IB[[1]] <- c( 0.2 , 0.2 , 0.2 , 0.2 , 0.2 )
IB[[2]] <- c( 0.8 , 0.1 , 0.05 , 0.025 , 0.025 )
IB[[3]] <- c( 0.05 , 0.15 , 0.7 , 0.05 , 0.05 )
sapply( IB , H )
```

which give the following output,

```
[1] 1.6094379 0.7430039 0.9836003
```

The first island has the largest entropy, followed by the third, and then the second in last place. Why is this? Entropy is a measure of the evenness of a distribution. The first islands has the most even distribution of birbs. This means you wouldn't be very surprised by any particular birb. The second island, in contrast, has a very uneven distribution of birbs. If you saw any birb other than the first species, it would be surprising. Now we need K-L distance, so let's write a function for it:

```
DKL <- function(p,q) sum( p*(log(p)-log(q)) )
```

This is the distance from q to p, regarding p as true and q as the model. Now to use each island as a model of the others, we need to consider the different ordered pairings. I'll just make a matrix and loop over rows and columns:

```
Dm <- matrix(NA, nrow=3, ncol=3)
for (i in 1:3) for (j in 1:3)
    Dm[i,j] <- DKL(IB[[j]] , IB[[i]])
round(Dm, 2)
```

with the following output,

```
     [,1] [,2] [,3]
[1,] 0.00 0.87 0.63
[2,] 0.97 0.00 1.84
[3,] 0.64 2.01 0.00
```

The way to read this is each row as a model and each column as a true distribution. So the first island, the first row, has the smaller distances to the other islands. This makes sense, since it has the highest entropy. Why does that give it a shorter distance to the other

islands? Because it is less surprised by the other islands, due to its high entropy.

SOLUTION 2. I won't repeat the models here. They are in the text. Model `m6.9` contains both marriage status and age. Model `m6.10` contains only age. Model `m6.9` produces a confounded inference about the relationship between age and happiness, due to opening a collider path. To compare these models using WAIC:

```
compare( m6.9 , m6.10 )
```

which gives the following output,

```
        WAIC pWAIC dWAIC weight    SE  dSE
m6.9  2714.0   3.7   0.0      1 37.54   NA
m6.10 3101.9   2.3 387.9      0 27.74 35.4
```

The model that produces the invalid inference, `m6.9`, is expected to predict much better. And it would. This is because the collider path does convey actual association. We simply end up mistaken about the causal inference. We should not use WAIC (or LOO) to choose among models, unless we have some clear sense of the causal model. These criteria will happily favor confounded models!