Lecture #2

Evan Kiefl

1/10/17

Note: *These lecture notes are still rough, and have only have been mildly proofread.*

Likelihood

Motivation

Suppose we have 2 models to explain some data. Which model is better?

Lecture Summary

The concept of a likelihood is defined for a model when presented with data. Suppose that you flip a coin with unknown bias 10 times and record the outcome as either heads or tails. Let $\underline{x} = (0, 0, 1, 0, 0, 0, 0, 0, 1, 1)$ be your data, where 0 corresponds to tails and 1 corresponds to heads. Suppose your model is that the coin is designed to give heads with probability p, and call this model M_p. The likelihood for M_p is defined as

$$L(M_p) = \Pr(\underline{x}|M_p)$$

Assuming the flips are all independent,

$$L(M) = \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i}$$

Let's write a function that calculates $L(M_p)$ when given $\underline{x}$ and p, and then calculate the likelihood for two different models, M_{p_1} and M_{p_2}, where $p_1 = 0.2$ and $p_2 = 0.7$.

```
L <- function(x,p) {
    prod(p^x * (1-p)^(1-x))
}

# x is the sequence of observations (our data)
x <- c(0,0,1,0,0,0,0,0,1,1)
# p1 is probability of heads
p1 = 0.2
p2 = 0.7

Lp1 <- L(x,p1)
Lp2 <- L(x,p2)

print(c(L_p1=Lp1, L_p2=Lp2))
```

```
##        L_p1          L_p2
## 0.0016777216 0.0000750141
```

We may have an intuition that the likelihood for M_{p_1} ($p_1 = 0.2$) will be higher for M_{p_1} ($p_2 = 0.7$), since our dataset only contains 3 heads. We can indeed see that this is the case, and so we favour M_{p_1} over M_{p_2}. We've demonstrated that likelihoods can be used to compare two models and make an informed decision about which one better describes the data. We can further quantify how much *more* likely M_{p_1} by calculating the likelihood ratio:

$$L(M_{p_1}, M_{p_2}) = \frac{L(M_{p_1})}{L(M_{p_2})}$$

For our example, we get $L(M_{p_1}, M_{p_2}) \approx 22$, which means M_{p_1} is favored over M_{p_2} by a factor of 22. This is great, but how impressed should we be with this number? In other words, how confident should we be that the data was indeed generated by model M_{p_1} instead of M_{p_2}? See http://stephens999.github.io/fiveMinuteStats/ LR_and_BF.html for a continuation of this idea.