Binomial distribution, Bayesian updating

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 13, Fall 2018

Wednesday March 6

Plan

- Probability and random variables
- Prior probability
- Binomial distribution
- Bayesian updating

Determining proportion of surface water on Earth

- Experiment: determine the proportion of Earth that is covered by water (p) (our unknown).
- We toss a mini globe in the air, catch it, and record whether our index finger is touching land (L) or water (W) on every toss.
- Outcomes of experiment are complementary: Either W or L.
- We generate data sequence: (e.g. W, L, L, W, W, W, L, ...)
- Probability of W: p
- Probability of L: $1 - p$ (why?)
- Tosses independent

Binomial distribution

- The binomial distribution does the job of assigning probabilities to the random variable we're interested in estimating: Proportion of surface water on Earth (p)
  - Binomial distribution tells us plausibility of each conjecture about p's value.
- It takes in 3 inputs
  - The probability of seeing W in a trial: p
  - Number of trials we perform: n
  - Number of Ws we observe in the n trials: w
- **We know** n and w as data (we're in charge of the experiment!) and **we don't know** p[1].

---

[1] We could fix and estimate other combinations (e.g. fix p and n, and estimate w, like in a fair coin toss experiment).

It has a scary formula

$$f(w|n,p) = \frac{n!}{w!(n-w)!} p^w (1-p)^{n-w}$$

- Remember: It is just a FUNCTION: assigning probabilities to our random variable!
- The notation $f(w|n,p)$ means that the binomial distribution returns for us the plausibility of observing w Water outcomes, GIVEN (|) n total globe tosses and the Earth being $100p$ % water.
- It is our job to give it hard numbers for w, n, p to get a probability output.
- If we need to give it all w, n, p and we don't know p, what do we plug in for p?
- We know p ranges from 0 to 1. Let's sample that interval at regularly spaced points. The binomial probability will clue us to the most likely value for p.

Bayesian updating

As new data comes in from our tossing experiment, we can update our best guess of what p is using **Bayesian updating.**

1. Sample some number of points for p (e.g. 20).
2. Plug them into the binomial distribution, as well as how many tosses we've done (n) and w Water outcomes we've observed.
3. Multiply by the probability distribution that describes our best guess on the previous toss (prior).
4. Normalize
5. Do another trial, and repeat

Let's code this up in Python and plot the results after each trial