Homework 5

Edward Trout

March 6, 2019

Question 1: For Bayesian analysis, we will be using the software package Stan. Stan works using a physics-based algorithm to sample from the posterior distributions.

A) Install the following R packages on your computer: rstan, rstanarm, shinystan

```
install.packages(c("rstan","rstanarm","shinystan"))

library(rstan)
library(rstanarm)
library(shinystan)
```

B) Search the Stan forums (http://discourse.mc-stan.org/) for an example of Stan code that is relevant to your project. Below, include a web link to the forum discussion and write a brief description of how you might adapt the discussed model for your own purposes.

As my project deals with the evaluation of occupancy given differing environmental and anthropogenic covariates, I investigated the occupancy modelling of this forumbase. I found one discussion of an example occupancy model from the stan manual involivng butterflies.

https://discourse.mc-stan.org/t/occupancy-model-error-with-k-value-number-of-visits-solved/5411

The forum discussion itself was about troubleshooting the run-through of this code, and since it is an example to work one's way into becoming familiar with stan, I believe this will be quite useful for me. The example itself is found at:

https://mc-stan.org/users/documentation/case-studies/dorazio-royle-occupancy.html

The main transormation or adaptation I can anticipate at this point is to change the visits parameter to ameliorate with my cameras. I can assume that 'visits' will be akin to 'days' for the cameras, so there will be a higher sample count than these butterfly visits.

Question 2: Read the following texts (both texts are available in github) and answer the associated question with approximately one paragraph of text:

A) Dall et al 2005, *Information and its use by animals in evolutionary ecology*

Question: Are foraging animals Bayesians? Why or why not?

The paper presented here does indeed make a compelling argument that foraging animals are indeed bayesians. On a theoretical basis foraging animals are informed by priors- either genetically from countless generations of revised distributions or experientially from constant practice throughout an individual's life. These priors then are used to drive a sampling process that accrues data of found forage given the quality of the forage area. These processes are repeated over and over again, all leading to the posterior probability of an area being quality for forage given the forage found there.

B) *The Garden of Forking Paths* by Jorge Luis Borges 1941.

Question: Why has the phrase "garden of forking paths" become so important and widespread in modern statistics?

In this piece by Jorge Borges, the idea of an unlimited number of paths and recursive histories is represented multiple times in phyiscal motifs the characters navigate, the narrative structure of the short story, and

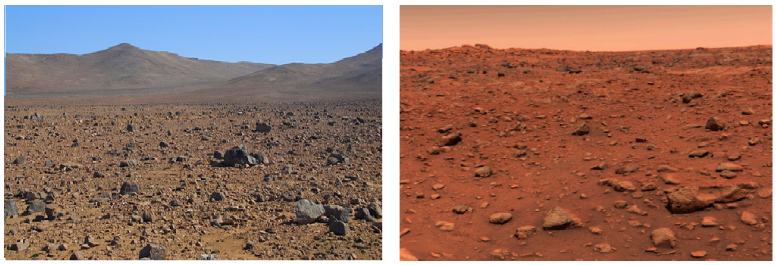

Figure 1: The Atacama Desert (left) and the Chryse Planitia (right)

in a diagetic novel titled the same as the short story. While all filled with the fascinating and fantastic, these themes have relevant and important implications for the very idea of statistics. It is the statiticians imperative to try and master these boundless paths, these alternate histories- alternate causalities, in order to capture and relate the mysteries of the universe. Upon observing a natural phenomenon, the central tenet of the *Garden of Forking Paths* must be adhered to. What are the cascading series of causalities that led to that phenomenon? This idea can be seen very explicitly in the basic Bayesian theorem. Given what we know about the world and given what we observe, what are the possibilities that support our hypothesis?

Question 3: Earth is 70% covered in water, while Mars is 100% land. A randomly selected satellite pixel from one of the planets was classified as "land." Assuming the satellite pixel was equally likely to be from Earth or Mars, show that the posterior probability that the pixel was from Earth, conditional on seeing land [P(Earth|Land)], is 0.23.

The basic Bayes equation states that a **posterior probabilty** $P(H|D)$ is equal to the **likelihood** $(P(D|H))$ of seeing data multiplied by the **prior probability** $P(H)$, all divided by the **average sum Likelihood of data** $P(D)$. In the text of mathematical formulae:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$$

In this given question, our **posterior** is the likelihood of a pixel being from Earth when it is classified as "land" $P(Earth|Land)$. This is stated in the question, and is the result we would be looking for if we knew a pixel was land but did not know the planet of origin (*e.g.* the Atacama Desert in Chile does suspiciously resemble Mars'Golden Plain).

The given value of the **posterior** in this question is .23, and we will be showing how this is true.

The **likelihood** of this problem is the likelihood of finding a land pixel on Earth, given as $P(Land|Earth)$. We can easily evaluate this because we know the probability of any given pixel on Earth being water- 70% or .70. This means that the probability of any given pixel on Earth being land is $1 - .70 = .30$.

The **prior** of this problem is anything we know ahead of time about any given pixel being from either Earth or Mars, or the probability that any given pixel is from Earth $P(Earth)$. We are told this, that it is equally likely to have a pixel from Earth or Mars. The probability therefore is $1/2 = .50$.

Finally the **average sum likelihood of data** in this problem is the probability of any given pixel being land $P(Land)$. We can evaluate this using what we know of the probability of finding land on either Earth or Mars. Earth is .30, Mars 1. The probability of land on Earth *OR* Mars is $\frac{.30+1}{2} = .65$. We divide by two because the sum of all probabilities for a given event must be 1, but this also makes sense with the idea of *average* sum likelihood.

Figure 2: look at her, she's perfect

Now we can compute our equation:

$$P(Land|Earth) = \frac{P(Earth|Land) \times P(Earth)}{P(Land)}$$

With our values:

$$.23 = \frac{.30 \times .50}{.65}$$

And this comes out to be true!

Question 4: Suppose there are two species of panda bear. Both are equally common in the wild and live in the same places. They look exactly alike and eat the same food, and there is yet no genetic assay capable of telling them apart. They differ however in their family sizes. Species A gives birth to twins 10% of the time, otherwise birthing a single infant. Species B births twins 20% of the time, otherwise birthing singleton infants. Assume these numbers are known with certainty, from many years of field research.

Now suppose you are managing a captive panda breeding program. You have a new female panda of unknown species, and she has just given birth to twins. What is the probability that her next birth will also be twins?

A) Compute the probability that the panda we have is from species A, assuming we have only observed the first birth and that it was twins.

Alright, before this becomes overwhelming, lets look at that cute picture of a panda.

Adorable. Now, lets define our Bayesian arguments! Our **posterior** is the probability of this being species A given it gave birth to twins, or $P(A|twins)$. Our **data likelihood** is the probability of species A having twins $P(twins|A)$. This we know from research is 10% of the time, or .10. Our **prior** is the probability that this turtle is species A without any knowledge of its offspring $P(A)$. This we know from general species information is .50, it is equally likely to find either species A or B. Finally our **average sum likelihood of data** is the probability of twins happening regardless of species $P(twins)$, which is $\frac{.10+.20}{2} = .15$.

Our equation then is

$$P(A|twins) = \frac{.10 \times .50}{.15}$$

giving us a probability that this individual is species A is .33 or 33%.

B) Suppose the same panda mother has a second birth and that it is not twins, but a singleton infant. Compute the posterior probability that this panda is species A.

In this problem, we are evaluating a new **posterior**- $P(A|singletonAFTERtwins)$. This successional probability of two conescutive independent events will leverage the simple rules of probability, that the combined probability of two independent events is the product of the probabilities of those two events. Thus our our **data likelihood** is then $P(singletonAFTERtwins|A) = P(twins|A) \times P(singleton|A)$. We can compute this as $P(singletonAFTERtwins|A) = .1 \times .9 = .09$. Our **prior** will remain at $P(A) = .5$. Our **average sum likelihood of data** is now $P(singletonAFTERtwins) = P(singletonAFTERtwins|A) + P(singletonAFTERtwins|B)$, which here is $P(singletonAFTERtwins) = \frac{(.1\times.9)+(.2\times.8)}{2}$. The total equation is then

$$P(singletonAFTERtwins|A) = \frac{.09 \times .5}{.125}$$

and results in .36 or 36%.

The added information suggests that this individual is 3% more likely to be species A than we considered before. Because it is so likely that either species has singleton infants, a singleton birth doesn't inform much differently for one species over another.

C) A common boast of Bayesian statisticians is that Bayesian inference makes it easy to use all of the data, even if the data are of different types. So suppose now that a veterinarian comes along who has a new genetic test that she claims can identify the species of our mother panda. But the test, like all tests, is imperfect. This is the information you have about the test:

a. The probability it correctly identifies a species A panda is 0.8

b. The probability it correctly identifies a species B panda is 0.65

The vet administers the test to your panda and tells you that the test is positive for species A. First ignore your previous information from the births and compute the posterior probability that your panda is species A. Then redo your calculation, now using the birth data as well.

Well well. If you need a glance at that panda pic again, go ahead. But now let's get started. We are ignoring all previous birth information, so our **prior** is back at $P(A) = .50$. Our new **posterior** is $P(A|positiveforA)$. Our **data likelihood** $P(positiveforA|A)$ is the identification success for the test, or .80. Finally, our **average sum likelihood of data** $P(positiveforA)$ is the combination of both of the "A positive" results: returning positive for A when the species indeed is A (.80), and returning positive for A when the species is actually B (1 - .65 = .35) $\frac{.80+.35}{2} = .575$.

The Bayes equation is therefore

$$P(A|positiveforA) = \frac{.80 \times .50}{.575}$$

And the probability that this species is species A is .696 or 69.6%.

If we are to include the birth data from before, our **prior** information is now different. We know from that separate analysis that this individual has a 36% chance of being species A $P(A) = .36$. All of our other arguments remain the same giving

$$P(A|positiveforA) = \frac{.8 \times .36}{.575}$$

and this returns .501 or 50.1%.

So in total with both the birth data and the test data, we are still 50-50 sure that this panda is species A. This will continue to be a cryptic tribe, for the time being...