Statistical Rethinking

notes from book and youtube videos given by McElreath

Can I come up with a question to answer, maybe stellar formation?

Fundamentally, we're trying to 'construct a posterior' based on data. The prior is the posterior with no data.

The notion is to create generative models from DAG's or process models. This can generate 'dummy data'. Then we hope to create statistical models that can analyse the synthetic data. Then you might provide it with real data.

Parameter and estimand are equivalent in my mind, I'm pretty sure. Estimand is Θ.

How does McElreath relate Bayesian statistics?

He refutes the idea of a sample size, the posterior distribution embodies the sample size.

- Posterior distribution is the *whole* mathematical object, you can try and summarise it. For instance, with intervals, there's nothing special about the interval.

Tsitsiklis says that a fundamental disagreement in inference is what the ultimate mathematical object we're trying to arrive at is. In the classical approach, the quantity we're looking for is a constant values, we don't know it, but if we did it would be a point. In the Bayesian approach this object should always be modelled as random variable, a distribution across values. It doesn't mean nature is random but my subjective experience is it being an rv.

It seems with McElreath's approach we're combining the model and the variable inference approach our Θ is some set of values that define a generative model for data we have.

This generative model can then, in theory generate new data points.

What are the components of Bayes Theorem?

$$P(\theta|\bar{y}) = \frac{P(\bar{y}|\theta)P(\theta)}{P(\bar{y})}$$

Bayes theorem operates withing a probability model or space.

In this context, the unobservable θ, is a conjecture and something that can be inferred from a observation vector $\bar{y}$.

We need * Probability for the event of the observation vector given our conjecture. * Probability of the observation vector.

Must have either:

Joint probability of the quantity of interest and the evidence/measurement.

What do we need to know?

- The total probability of evidence.
- The base rate (prior) probability of the hypothesis, this is what we're trying to update.
- The joint probability of the evidence and the hypothesis.

Whats an alternative formulation of Bayes Rule using an example?

The common formulation is as such:

If vampires are very rare in the population say P(V) = 0.001 but we have a test that says with a 95% true positive rate if someone is a vampire. We select a random person from the population, they test positive for vampirism, what is the probability they're actually a vampire?

$$P(V|P) = \frac{P(V)P(P|V)}{P(P)} = \frac{0.001 * 0.95}{(1 - 0.001) * 0.05 + 0.001 * 0.95}$$

We have to normalise by all the ways you could see the data (get a positive test). In this case, you can get a positive when you're actually a vampire (0.001*0.95)

Randomness as a property of information not of the real world. I think a good example of this is the Monty Hall problem, we search for some ontological basis for updating our beliefs but it doesn't exists, its purely an informational change.

It seems useful to think of a distribution as a vector of discrete values, even if its continuous, just in terms of when computation is manipulating them.

What is the geocentric notion McElreath's trying to get across?

This is a model of prediction without explanation. Mechanistically wrong.

How does grid approximation compare to a Laplace approximation?

Grid approximate discretises the parameter theta.

Lecture 1

- Learning statistics to use it for scientific questions.
- Causal inference implies some predictive aspect to the model.
- Causal imputation, be able to observe counterfactuals, or 'what ifs'?
- T tests and general statistic often used to test null hypothesis, McElreath says that research science is about more than this kind of quality control, useful in industry (t test to get same experience) and experimental science.
- These industrial controlled settings are not the norm, the ability to do experimental interventions is limited. Call these 'observational'.
- Bayesian data analysis allows you to take the assumptions from your generative model and confront them with the least fuss (?) ~39 min.
- Bayesian models are generative, so it aligns with the models underlying how we're approaching answering scientific questions.
- Drawing the Bayesian Owl
  - Theoretical estimand, what is it we're trying to predict or answer a question about.
  - Causal model, develop some sort of causal model that eventually, should be generative.
  - Use the previous steps to build a statistical model.
- Dag's: Transparent scientific assumptions to justify effort, expose to critique and connect theories to golems.
- Golems: Statistical models or devices (brainless).

Lecture 3

workflow, from a scientific question, to the development of a causal model and from there to a Bayesian estimator

- Trying to make that distinction between a statistical model, like linear regression and a causal model. So we project some causal model on to the 'geocentric' model that is linear regression (really accurate but it's causality doesn't exist).
- There are many more ways for a sequence of coin tosses to put you on the half way line than away from it.
- The coin toss reduces likelihood that you'll get a sequence of right or left movements.

Gaussian is a model with very little assumptions (mean and variance).

(1) State a clear *question* Describe the association between adult height and weight

(2) Sketch your causal assumptions. Causal model: weight is some function of height.
(3) Use the sketch to define a *generative* model Assume that they effect each other with no mechanism.
(4) Use the generative model to build an *estimator* Want to estimate how the average weight changes with height.

Conceptually useful to defined unobserved things that might affect height (eg causality).

Generative model starts out as $W = \beta H + U$(unobserved stuff).

Estimator: $E(W_i|H_i) = \sigma + \beta H_i$.

- When you plot out your assumptions, at about the 55 min mark, it's interesting to maybe see how wild your assumptions are!
- There are no correct priors, just scientifically justifiable ones.

Chapter 1

- I struggled with the example about neutral theory of evolution as a hypothesis
- It seems that he's arguing against the falsifying of null hypotheses and for creating multiple non-null models of the natural phenomena. This is like creating explanations that can be falsified (Deutsch?)
  > - In order to challenge these process models with evidence, they have to be made into statistical models. This usually means deriving the expected frequency distribution of some quantity-a "statistic"- in the model
- Change your explanation to fit the process models and the statistic models they produce in accordance with the observed data
  > - Bayesian inference is no more than counting the number of ways things can happen, according to our assumptions. Things that can happen more ways are more plausible
- Frequentist approach struggles when there is no sampling invariance. The example used is Galileo looking at a blurred Saturn with its rings, no amount of re sampling will resolve the uncertainty present in the technology. The uncertainty is not a function of the repeated measurements
- With the Bayesian method this uncertainty in the information can be modelled.
  > - Bayesian golems treat "randomness" as a property of information, not of the world
- Nothing in the real world is actually random it just lacks information (hmmm?)

> - We just use randomness to describe our uncertainty in the face of incomplete knowledge

Chapter 2

- A *conjecture* is usually called a *parameter*.

I kind of like this example of the water on the globe. It's kind of what you're trying to always do with Bayesian Analysis (I think) which is postulate a state of the world.

Basic design loop: * Data story: Narrating how the data might arise. Can be descriptive and causal. Almost as if the data is a character, and you're trying to find motivations of why it exists or the reason it is the way it is.

"Bayesian data analysis usually means producing a story for how the data came to be".

- A Bayesian model begins with a set of plausibilities for each conjecture (priors).

Components of Model

- The no. of ways each conjecture could produce the observation.
- The accumulated no. of ways each conjecture could produce the entire data.

Unobserved variables (in our case, the proportion of water) are usually called *parameters*. We can then have observed variables, like what we draw from the bag of marbles, or what comes up under our finger tossing the globe (W or L).

Assign plausibility of p with the data (observables). We were able to defined the 'state of the world' through one variable p in the marble case (that is, the proportion that were blue).

The story as McElreath puts it is that we have to events, W and L. Nothing else can happen. We are given a string of 9 events (in this examples). Out of all the possible worlds where 9 events occur, with our parameter p defining what is the case, what is the plausibility of the string of 9 events we have.

A binomial distribution is counting the paths for you. For a given proportion of water to land, it's saying how many So we have some variable p, that constrains our sample space. On determining a new path, p is ever present. We have W, L which we might consider the data. The observables. Given that

- Rethinking datum/parameter? Data is normally considered 'known' and 'parameters' unknown.

2.4 Or story is we want to know the plausibility of p given some observable W out of N tosses.

The binomial distribution gives us a set of plausibilities for P(W,L|p). We just want this for every p.

The initial goal was to determine which conjecture, out of a set of conjectures was the most plausible given some data. In the marble example, we had 4 possible conjectures. Moving on to the globe example, the conjectures are all the possible states of the world (literally), this state is defined by the proportion of water to land.

Plausibility for a given conjecture is proportional to the plausibility of the data given the state of the world is our conjecture times the plausibility of that world being the case (prior).

This prior can also be thought of as the prior number of paths (for some previous data say). So it's just counting paths.

Chapter 2

- Small world vs large world. The example of Behaim's globe is used (it doesn't have the Americas in it). While the small world model is internally consistent is doesn't represent reality fully. It's this interplay between your small world (model) and reality thats important.
- Follows on here from the Bayesian inference explanation above. Looking at the garden of forking paths in which alternative events are cultivated as we learn what did happen some of these alternatives are pruned. In the end what remains, is what is logically consistent with our knowledge
- Counting the possible paths then becomes a multiplication of the possible paths on each ring (in the example)
- Marble example (I think p here is just a way of numerically describing the no. of blue marbles)
  - A conjectured proportion of blue marbles **p** is usually called the **parameter** value. It's just a way of indexing possible explanations of the data.
  - The relative number of ways that a value **p** can produce the data is usually called the **likelihood**. It is derived by enumerating all the possible data sequences that could have happened and then eliminating those sequences inconsistent with the data
  - The prior plausibility of any specific **p** is the **prior probability**
  - The new, updated plausibility of any specific **p** is the **posterior probability**

2.2 Building the model

- Creating a data story. A narrative of why we are getting the observations. Viewed as important as it makes you think of the variables you really need to consider, get a bit more exact about chain of events (creating something hard to vary?)

- walks through Bayesian updating, the amount of evidence we have is embodied in the plausibility (straight line at the start vs complex curve at the end). The final figure is normally shown but its important to know that it is just an iterative development from the first figure.
- Some tips given for evaluation, they seem rather abstract at the moment though with my current knowledge
- Now we look at mapping some of the concepts from the previous section to build up the model

Likelihood function (1) the number of ways each conjecture could produce an observation

- Because both outcomes (W and L) are equally likely, and independent we look at all the ways our sample size of 9 (n) could appear to us.
- The binomial distribution calculates the relative no. of ways to get six W's with 9 tosses holding p at 0.5
- Looking at the parameters to the binomial function p, n and w they can each represent different conjectures once we can tell the likelihood and what has been observed
- In the sciences, the prior is considered part of the model, there is no reason not to interrogate it like other assumptions.
- Bayesian estimate is always a distribution over the parameters
- Posterior is proportional to the product of the prior and the likelihood
    - Count up all the ways you could see the data and multiply by the priors (look at table for marbles)
- Grid approximation builds up from the marble example. With just 3 possible values for water (0, 0.5 and 1) 0.5 wins out with in 3 tosses. If we bump up the possible values to 20 though we get a more accurate display of possible values (posteriors)

Chapter 3

- Whenever the condition of interest is very rare, having a test that finds all the true cases is still no guarantee that a positive result carries much information at all.
- Interesting note box here. Why statistics can't save bad science. Suppose the probability of a positive finding and a false positive rate thats very low. If the probability of the prior, that is the probability of any hypotheses you posit in general being true is low, the best you could probably do is 0.5 (in terms of the posterior that the finding indicates the hypothesis is true). The lesson here being that no amount of accurate instrumentation can account for bad hypothesis (or explanations)

> - 95% is a small world number.. true in the model's logical world

  - On interpreting confidence intervals

- Loss function. If you were to make a bet on what the correct parameter value is. Where to cost is proportional to your distance from the 'correct' answer
  > - Given a realized observation, the likelihood function says how plausible the observation is. And given only the parameters, the likelihood defines a distribution of possible observations that we can sample from, to simulate observation

- Posterior predictive distribution

  - The top graph is the posterior distribution, for each parameter we are running is through a binomial distribution of 9 tosses as if the correct proportion p is that chosen parameter (so set p to 0.1 and run a simulation with p as 0.1 and see what posterior distribution you get as a result from 9 tosses). The initial posteriors are multiplied then by each sampling distribution and the final predictive distribution is shown
  - Passing the uncertainty of all parameters down is important so that the model does not appear more confident than it in the prediction

- Highlights different ways to analysis the model through alternative ways of a predictive distribution looking to see if the globe could be bias by the amount of times its switched, as an example

3.3

- Bayesian models are always *generative*.
- Generating implied observations from a model is useful

  - We can also sample from the prior, seeing what the model expects can tell us a lot about our assumptions, or the implications of our prior.
  - For testing, running through known data and checking that our model simulates expected observations.

- From chapter 2, we developed a model that built up a likelihood function based on observed data. We can now use this likelihood function to think about what we might observe next.