Day 12 - Intro Bayesian

David Carlson

December 22, 2021

Bayes' Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$
$$= \frac{p(A\&B)}{p(B)}$$

- $p(A|B)$ is a conditional probability: The probability of event (or distribution) A occurring given that B (an event or distribution) is true
- $p(B|A)$ is also conditional, just the reverse
- $p(A)$ and $p(B)$ are marginal probabilities
- $p(A)$, just as a preview, can be known or treated as a prior when we move on to statistics
- $p(B)$, on the other hand, is the marginal likelihood, and is often ignored (sort of), as it acts as a normalizing constant (probability must sum or integrate to one)
- A and B must be different (though not independent, clearly) events
- When we put this in statistical terms, we want to know the distribution of our parameter space, which we denote θ, given the observed data, $\mathcal{D}$
- This leads us to the formulation of Bayes' theorem for statistics: $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$
- So now the $p(\mathcal{D}|\theta)$ should be familiar from frequentist statistics and p-value testing (probability of observing the data given the null hypothesis is true)
- But this is not what we actually care about
- We want to make inferences on the parameter space given the data we observe
- So we are after $p(\theta|\mathcal{D})$, which is theoretically much more sensible
- The term $p(\theta)$ is the prior we place on the parameter space (can be flat, or uninformative, or informative)
- The marginal likelihood $p(\mathcal{D})$ is mostly a nuisance
- We can determine the marginal likelihood analytically, or through simulation, but what we really care about is maximizing the conditional probability of the parameter space
- Before we get more into stats, let's look at an example of the theorem in practice, which is a theorem about probability, so we will focus on events to make it more intuitive

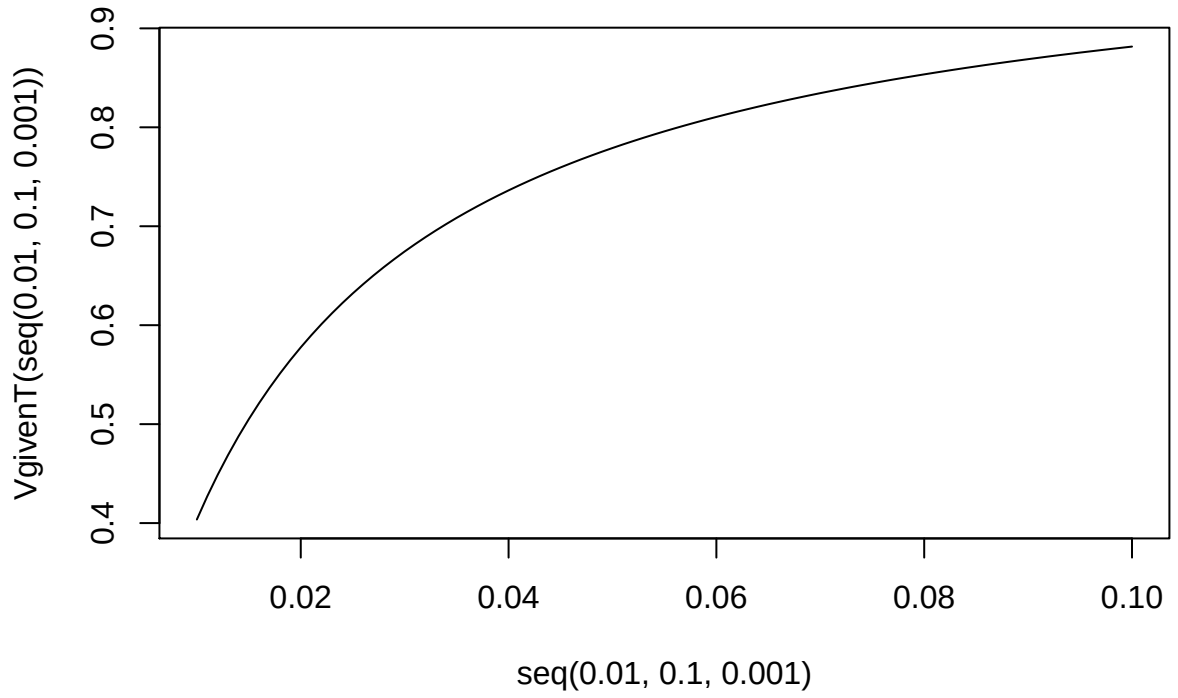
Covid-19

- What we are really interested in is the probability of somebody being infected with the virus after getting back a positive (or negative) test result $p(V|T) = \frac{p(V \cap T)}{p(T)} = \frac{p(T|V)p(V)}{p(T)}$
- Now we are conditioning on positively tested people
- We look at all the people who were positively tested
- We still need the probability of being tested positive

$$
\begin{aligned}
p(T) &= p(T \cap V) + p(T \cap V^C) \\
p(T \cap V) &= p(T|V)p(V) \\
p(T \cap V^C) &= p(T|V^C)p(V^C) \\
&= (1 - p(T^C|V^C))p(V^C) \\
p(T) &= p(T \cap V) + p(T \cap V^C) \\
&= p(T|V)p(V) + (1 - p(T^C|V^C))p(V^C) \\
p(V|T) &= \frac{p(T|V)p(V)}{p(T|V)p(V) + (1 - p(T^C|V^C))(1 - p(V))}
\end{aligned}
$$

- Now it only depends on sensitivity ($p(T|V)$), specificity ($p(T^C|V^C)$), and overall infection rate ($p(V)$)
- Sensitivity is roughly 0.67 for the most sophisticated tests
- Specificity is roughly 0.99
- Let's look at what happens when our priors change- Priors will change based on location, risk factors, etc.

```
VgivenT = function(prev) .67*prev/(.67*prev + .01*(1 - prev))
prevs = seq(.01, .1, .001)
plot(seq(.01, .1, .001), VgivenT(seq(.01, .1, .001)), type = 'l')
```



Moving from Bayes' Theorem to Statistics Applications

- A very common misinterpretation is that in frequentist, the data is random, and the parameter is fixed, while in Bayesian the data is fixed, and the parameters are random
- This is a fairly accurate representation of frequentist stats, but is wrong in its interpretation in Bayesian
- The flaw comes with what is 'given'
- In proper Bayesian, *everything* is random
- What we observe is a random realization of an underlying distribution, and even if we observe the entire population of interest, the observational values are still only a representation of the *distribution*

- In other words, randomness is inherent
- As an example, genetic distributions of the entire global human population still involve some random distributions that 'dictated' that observation, but the realizations are random, they are perturbations of some underlying distribution
- Bayesian stats is very much obsessed with randomness, which is arguably theoretically much more sensible
- Much more importantly in application is that we get the probability distribution of the parameter space
- Instead of seeking out the probability of the data given the null, we get the conditional probability of the parameter space, given the data (though again, not assumed fixed)
- We can therefore make inferences about the distribution of the parameter space, rather than relying on comparison to the null
- Claims of, for example, 95% confidence, are actually 95% of the mass of the estimated parameter space distribution in Bayesian
- This is one reason we no longer call it a $confidence$ interval - it is a *credible interval*
- The given data or probabilities as mentioned before, as the most important example, the prior, is repreented as a distribution, not a fixed point
- So, let's say the prior belief is solely based on location, not risk factors
- We know risk factors also play a role, so even though there may be a point estimate of prevalency in a location, there is some distribution underlying the individual observations (people)
- This distribution can take any form, or be empirical
- Think of it as you don't know the risk factors (since that could be empirically difficult / impossible / unethical)
- In practice, we simply do not have the complete information, so we can specify an underlying distribution (informative), assume relatively flat distributions (uninformative), or use empirics (empirial Bayes)
- Unfortunately, in our discipline, unless you're doing a meta-analysis, we tend to use uninformative priors, but this is changing (of course the spear-headers are Americanists)
- What it also buys us, other than theoretically driven inferences that do not rely on incorrect interpretations of the CI etc., is huge flexibility
- We can, in theory, model any data-generating process (DGP)
- If you think there may be a correlation, for example geogrpahic contagion or temporal autocorrelation, you can code that up
- We just have to code the Bayesian model to allow for and take into account any aspect of the DGP that may be of interest or impactful to our inferences
- Also, as with RE vs. FE, the distributional assumptions 'borrow' information for other units of study
- This means that we have more statistical power, and can make much sounder inferences
- We are after generalizable claims; with weak priors we can say things about the distribution of causal effects or correlations
- So, we can make inferences about populations that *are not even studied*, and take into account the uncertainty about them

Bayesian Statistics

- We are interested in $p(\theta|\mathcal{D})$; that is, the probability distribution of our parameter space given the data
- Again, everything is random
- So, we model things in a linear model, e.g., as $y_i \sim \mathcal{N}(X_i\beta, \sigma^2)$
- Once AGAIN, everything is RANDOM
- The observables in X_i are randomly distributed realizations of a population of interest
- The population of interest, even if you had every person is the world, is still random with an underlying distribution
- Every new person will be another random realization
- We want to understand that distribution
- We model it very similarly in spirit to MLE, but much differently in practice and code

- Let's think about MCMC (Monte-Carlo Markov-Chain) for Bayesian inference
- We will get much more in the weeds next class
- For now, how it basically works is that there is a random walk in the parameter space
- If that walk provides a better estimate of the parameters that explain the data, keep it, if not, keep it with some probability
- We then look at the distribution of kept, or accepted, values of the parameter space
- This distribution in theory approximates the distribution of the parameter space given the population
- We can then infer evidence from the approximate parameter space distribution
- We can also look at the parameter space distribution, the 'effects,' and make much more precise claims about the effect at a distributional level
- This means no more Z-scores, t-statistics, etc., the parameter distribution we estimate with iterations is precisely the distribution we are after
- We can then look at the 0.95 bounds of the effect in a counter-intuitevly less distributional manner as compared to frequentist
- Unfortunately, the current state of the discipline wants these bounds of effect sizes
- We cannot plot the distribution because it is high-dimensional (usually)
- So we can look at features; e.g. the substantive implication section can be much, much more detailed and precise

Very Elementary Sampling

- Start with a random guess for the parameter space
- Check the probability of that parameter space
- Randomly walk to a new parameter space, and check that probability, and decide if you should retain that value of the parameter space
- If you reject it, go back to the original suggestion, and if not, move on to that guess
- Repeat
- In practice we sample one parameter at a time and sequentially move throguh them
- We then make the decision based on the joint distribution
- This is coherent, because the algorithm is constantly checking the parameter space
- We are basically randomly smapling with some decision criteria what best explains the data, but with randomness
- We end up with a distribution of the likely parameter space
- Because we start randomly, we designate the first half (in general) of the sample as burn-in, and the second half as sampling from the 'true' distribution
- We throw out the burn-in as it is uninformative
- We keep the sampling distribution as an appoximation to the true distribution
- Then we infer
- This is called Monte Carlo Markov Chain
- It is Monte Carlo because of the random walks
- It is a Markov chain because every iteration depends on the previous iteration, so are dependent, but the entirety of the chain is independent, because each iteration is only based on the previous iteration, and only dependent through the dependence of the last iteration
- We will in practice almost always use Hamiltonian MC
- HMC takes into account the geometry of the parameter space and the data, so is much, much faster, with each iteration slightly more expensive but the convergence to the space of interest requires much fewer iterations
- Imagine a hockey puck on ice sloped towards a maximum conditional probability
- If it is far away, it will slide faster, and if it is close, it will slide slower
- We then get an overall 'picture' of the distribution with much less sampling