Jeopardy Week 9

Why would we use MCMC? What is it good for?

# MCMC is a way to approximate the posterior distribution that are difficult to determine

Markov Chain Monte Carlo methods will allow us to approximate the posterior distribution of the parameter of interest.

What does MLE stand for?

Maximal Likelihood Estimate

What does MCMC stand for? Markov Chain Monte Carlo

What are the three part of MCMC

Monte Carlo

Markov Chain

Accept Reject Sampling (Metropolis Hasting Algorithm)

What does Monte Carlo Do?

What does Markov Chain do?

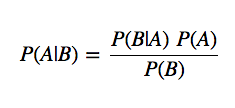
when we sample random numbers, we're going to rely on the previous random number to help randomly generate the next one

this is influenced by markov chain, which uses the point before it to generate the next number (remember the matrix multiplication we did on the board!)

What is conjugacy?

Conjugacy is where our prior and our posterior follow the same parametric form - the same type of distribution

What is Baye’s Theorem? Both in mathematical notation and in words.



Write out Bayes equation and explain what each element is. Include:

Here, we're representing the probability of flipping with:

Our **prior** belief (before observing flips) of the probability of flipping heads

The **likelihood** of the data we observe given the chance to flip heads

The **total probability** of observing that many heads in coin flips regardless of weighting

Posterior – a distribution of our parameter of interest given our prior and likelihood

How do we calculate likelihood?

Number of ‘successes’/total number of trials (n/k)

What is marginal probability?

The marginal probability of the data is the probability that our data are observed regardless of what model we choose or believe in. You divide the likelihood by this value to ensure that we are only talking about our model within the context of the data occurring.

P(B)

Is using Bayesian stats less scientific than using a frequentist perspective? Why or why not?

What are the three types of missingness?

Missing completely at random – where they’re missing entirely randomly

The data of interest is not systematically different between respondents and nonrespondents

Missing at random – some people may decide to just not respond to a specific question and leave it blank

I administer a survey that includes a question about income. Those who are female are less likely to respond to the question about income. § This type of missingness is called missing at random (conditional on sex). - Conditional on data we have observed, the data of interest is not systematically different between respondents and nonrespondents

Not missing at random – if a particular group of people are less likely to respond to a question on a survey (lower income participants may be less likely to respond to a question

I administer a survey that includes a question about income. Those who have lower incomes are less likely to respond to the question about income. § This type of missingness is called not missing at random. - The data of interest are systematically different for respondents and nonrespondents. - Whether or not an observation is missing depends on the value of the unobserved data itself!

What is the worst kind of missingness?

Not missing at random

Unit non response – no values from an observation are observed – they’re just entirely blank

Item nonresponse – some but not all values from an observation are observed (swiss cheese response)

Describe weight class adjustements when dealing with missing data:

Take a full sample (respondents and nonrespondents) and break the into strata based on known characteristics (age, sex, department, major, etc)

Reweight the responses: true proporation/ proportion of responses.

We have 2 departments of equal size, finance responds 25% of the time and HR responds 75% of the time, weighting would look like:

Finance = 0.5/0.25 = 2

HR = 0.5/0.75 = .666

With weight class adjustments, how does bias and variance change?

Decrease bias, but could increase variance

What is imputation?

Filling in missing values

What is deductive imputation?

If you have a null value in column of a row, you can use the surrounding information to deduce what that value might actually be (children example) using logical relationships to fill in missing values

Impute with mean median and mode: what are the pros and cons of this?

Easy to implement and undertand

Con: can distort a histogram, can underestimate variance in the data (may make it more biased)