Class 10: Introduction to Bayesian Statistics & Econometrics

MFE 402

Dan Yavorsky

Last Class

- Properties of ML Estimators
  - Not necessarily unbiased
  - Invariant to re-parameterization
  - Excellent asymptotically:
    - Consistent
    - Asymptotically Normal (and we derived their standard errors)
    - Asymptotically Efficient (achieve Cramer-Rao Lower Bound)
- Logit Example

Topics for Today

- Introduce Bayesian Statistics
- Learn the Lingo
- Simple Case: Conjugate Priors
- Common Case: Posterior Sampling

Introduction to the Bayesian Perspective

It's all about probability

In life, we ask questions, develop theories, and make predictions.

- How much does a particular drug affect a patient's condition?
- What can an average student earn after obtaining a MFE degree?
- Will the democrats win the next US presidential election?

But answering or testing them is not easy! Life is complicated, and it is often impossible to exactly isolate the parts of a system which we want to examine. Noise obfuscates signal.

Statistical inference is the logical framework which we can use to trial our beliefs about the noisy world against data. We formalize our beliefs in models of **probability**.

Frequentists

Classical / Frequentist perspective:

- Suppose a sample of data is the result of an infinite number of exactly repeated experiments
- The sample obtained is assumed to be the outcome of some probabilistic process
- Any conclusions drawn are based on the supposition the events occur with probabilities, which represent the long-run frequencies with which those events occur
- This probability actually exists, and is fixed for each event/experiment that we carry out
- Data are assumed to be random, resulting from sampling from a fixed and defined population distribution (ie, data generating process)

For example, a coin flip:

- $\mathbb{P}$(heads) is the proportion of heads observed in an infinite number of coin tosses
- A sample of coin flips (data) is generated as if part of that infinite series
- If there are 7 heads on 10 flips, this is the result of a slightly odd sample

Bayesian

Bayesian Perspective:

- Probability is a measure of certainty in a particular belief
- Different people can have different beliefs
- No repeatability is necessary (or necessary even to imagine)
- These subjective beliefs (probabilities) can be updated in light of new data
- Data are observed, and thus fixed

For example, a coin flip:

- $\mathbb{P}$(heads) is our belief about how likely we are to get a heads on the next flip
- The "parameter" here might be fixed or might be varying
- If there are 7 heads on 10 flips, we can update our belief about the probability of heads

Frequentist vs Bayesian Inference

Inference:

- We would like to know the "probability of a hypothesis given the data observed"
- When we choose a probability model to describe a situation, it enables to calculate the "probability of obtaining our data given our hypothesis being true"
- **This is the inverse/oppose of what we want!**

Frequentists do hypothesis tests:

- They directly use that inverse probability as evidence for/against a given hypothesis
- Assume hypothesis is true and calculate the probability of obtaining the observed data
- If that probability is small, then it is assumed unlikely that the hypothesis is true

Bayesians:

- Invert the inverse probability to get exactly what they want
  (ie, the "probability of a hypothesis given the data observed")
- Doing so requires a "prior"
  a probabilistic statement of beliefs in the hypothesis before we collect/analyze data

Bayesian Inference via Bayes' Rule

$$\text{Bayes' Rule:} \qquad \mathbb{P}(\theta|\text{data}) = \frac{\mathbb{P}(\text{data}|\theta) \times \mathbb{P}(\theta)}{\mathbb{P}(\text{data})}$$

The **Likelihood** is $\mathbb{P}(\text{data}|\theta)$.
Once we choose a statistical model, this is (usually) easily calculable for a particular θ.

The **Prior** is $\mathbb{P}(\theta)$.
It represents the researcher's pre-data beliefs about values of parameters.

The **Denominator** is $\mathbb{P}(\text{data}) = \int \mathbb{P}(\text{data}|\theta)\,\mathbb{P}(\theta)\,d\theta$.
It is fully determined by our choice of statistical model and prior.

The **Posterior** is $\mathbb{P}(\theta|\text{data})$.
This is the goal of Bayesian inference.

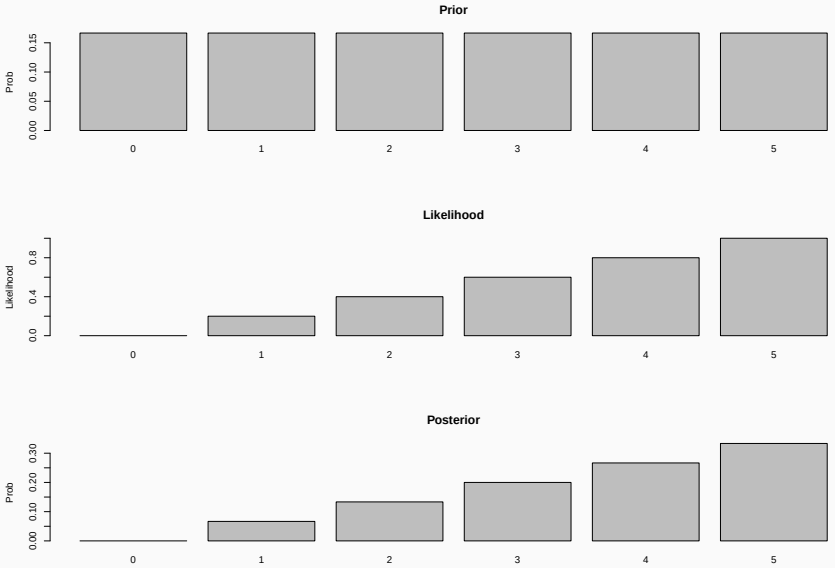
Fish Example: flat prior

Imagine a fish tank that you cannot see into, containing 5 fish, each of which is either red or white. We want to estimate the total number of red fish in the bowl after we pick out a single fish and find it to be red.

Num Red	Prior	Likelihood	Prior $\times$ Lik.	Posterior
0	1/6	0	0/30	0/15
1	1/6	1/5	1/30	1/15
2	1/6	2/5	2/30	2/15
3	1/6	3/5	3/30	3/15
4	1/6	4/5	4/30	4/15
5	1/6	1	5/30	5/15
Total	**1**	**3**	**1/2**	**1**

Notice how the posterior is a weighted average of the prior and the likelihood.

Fish Example: flat prior

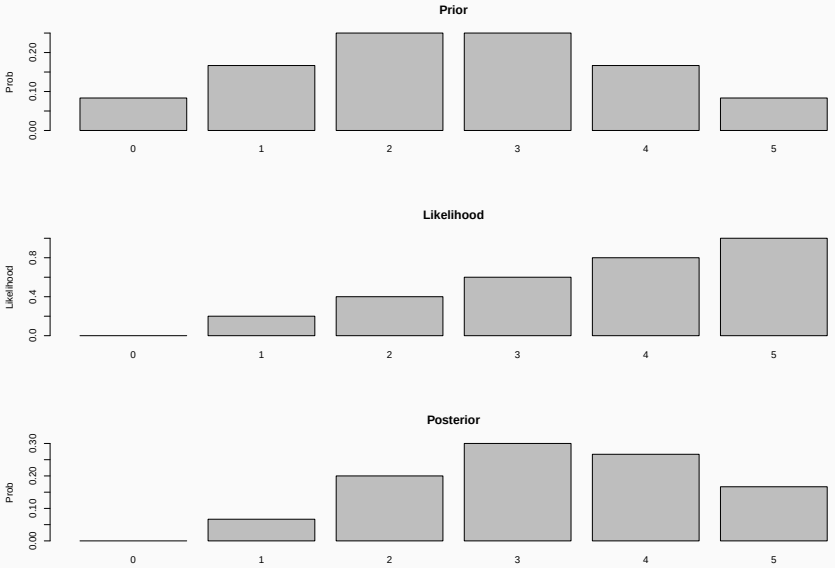


Fish Example: informative prior

Suppose instead that we believe that extreme numbers of red/white fish are less likely, and that a mix of red/white fish is more likely. Our prior accordingly captures this belief:

Num Red	Prior	Likelihood	Prior $\times$ Lik.	Posterior
0	1/12	0	0/60	0/30
1	2/12	1/5	2/60	2/30
2	3/12	2/5	6/60	6/30
3	3/12	3/5	9/60	9/30
4	2/12	4/5	8/60	8/30
5	1/12	1	5/60	5/30
Total	**1**	**3**	**1/2**	**1**

Fish Example: informative prior



The Prior

Practically:

- We need a way to apply Bayes' Rule so we must state a prior.
- Can use a "non-informative" (or flat) prior when the parameter space is discrete, but not when it's continuous

Philosophically, two interpretations:

1. The *subjective state of knowledge* interpretation is that we use a probability distribution to represent our uncertainty over a parameter's true value (assuming that there is such a thing as a true value for a parameter)
2. The *population* interpretation is that parameters themselves come from a distribution for each sample/observation, and our prior is our belief about that distribution

A comment about priors and subjectivity

A major argument *against* Bayesian statistics is that it is subjective due to its dependence on the researcher specifying their pre-experimental beliefs through the prior.

Bayesians argue that *all* analyses involve a degree of subjectivity, either explicitly or implicitly:

- The choice of statistical model (ie, the form of the likelihood)
- The choice of data to include (all variables in the dataset, outliers)
- The choice of specification (X, X^2, $\log(X)$, etc)
- The way in which models are (or are not) checked/tested
- The choice of 95% (for hypothesis testing by Frequentists)

Bayesians also defend their approach by saying

- They're more explicit about the subjective elements of their analyses
- The more data you have, the less "influence" is exerted by the prior

The Likelihood

This is the same object from our study of Maximum Likelihood Estimation.

Given the observed data, the likelihood is a function over the parameter space that indicates which values of the parameters most likely generated the observed data.

This is also where most of the statistical model resides.

In the fish example,

- the likelihood is the binomial distribution, viewed as a function of θ (the percent of red fish in the tank)
- $L(\theta) = p(y; \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$

The Denominator

A valid probability distribution is (1) non-negative, and (2) integrates to 1.

The numerator of Bayes' Rule is not a valid probability distribution:

- The likelihood is not a valid probability distribution (it doesn't integrate to 1).
- Similarly, multiplying it by the prior does not create a valid probability distribution.
- But the numerator is non-negative
- It just needs to be scaled so that it integrates to 1

The denominator provides that scaling.

Notice that the denominator is not a function of θ, and so it is just a normalizing constant. Thus, we can write Bayes' Rule as:

$$\mathbb{P}(\theta|\text{data}) \propto \mathbb{P}(\text{data}|\theta) \times \mathbb{P}(\theta)$$

The Posterior

With maximum likelihood estimation, the *whole* likelihood function is the "answer"

With Bayesian statistics, the *whole* posterior probability distribution is the "answer"

The posterior is the combination of our pre-data beliefs, updated with the observed data. It is a probability distribution over values of the parameters.

You can use the posterior to produce a point estimate, common choices are:

- The posterior mean: $\mathbb{E}[\theta|\text{data}] = \int \theta \, \mathbb{P}(\theta|\text{data}) \, d\theta$
- The posterior median
- The maximum a posteriori (MAP): ie, the highest point on the posterior

Frequentist Confidence Intervals vs Bayesian Credible Intervals

A Bayesian would say that Frequentist confidence intervals:

- are hard to understand – if I repeated this sampling and estimation process 100 times, 95/100 of those intervals would contain the true parameter value
- and can be meaningless – 5/100 of those intervals are just some range of numbers that doesn't contain the true parameter value

Bayesians construct credibility intervals.

- Take a section of the posterior distribution with bounds L and U that integrates to 95%
- There is a 95% chance that θ is between L and U

Bayesian Updating

There is a logical consistency to Bayesian inference as more data are collected:

- We begin with a prior $\mathbb{P}(\theta)$
- We observe data $\mathbb{P}(\text{data}|\theta)$
- We update our beliefs to the posterior $\mathbb{P}(\theta|\text{data})$
- The Posterior can then serve as a prior for additional data

For example, for the first observation of data, we have:

$$\mathbb{P}(\theta|y_1) \propto \mathbb{P}(y_1|\theta) \times \mathbb{P}(\theta)$$

Then for the second observation of data, we have:

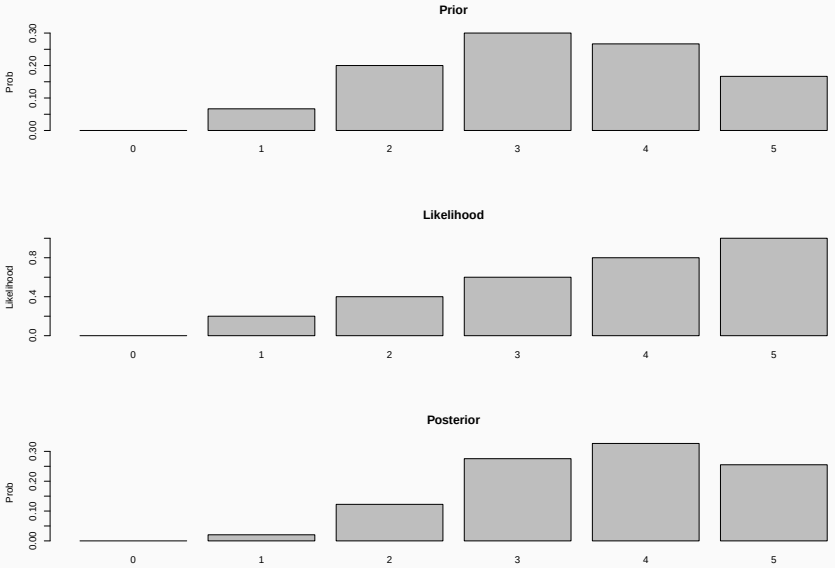
$$\begin{aligned}\mathbb{P}(\theta|y_2, y_1) &\propto \mathbb{P}(y_2, y_1|\theta) \times \mathbb{P}(\theta) \\ &= \mathbb{P}(y_2|\theta) \times \mathbb{P}(y_1|\theta) \times \mathbb{P}(\theta) \\ &= \mathbb{P}(y_2|\theta) \times \mathbb{P}(\theta|y_1)\end{aligned}$$

Fish Example: Bayesian Updating

Let's use the posterior from the previous example as our new prior. We return the red fish to the tank and select another fish. This second fish is also red.

Num Red	Prior	Likelihood	Prior $\times$ Lik.	Posterior
0	0/30	0	0/150	0/98
1	2/30	1/5	2/150	2/98
2	6/30	2/5	12/150	12/98
3	9/30	3/5	27/150	27/98
4	8/30	4/5	32/150	32/98
5	5/30	1	25/150	25/98
Total	**1**	**3**	**98/150**	**1**

Fish Example: Bayesian Updating

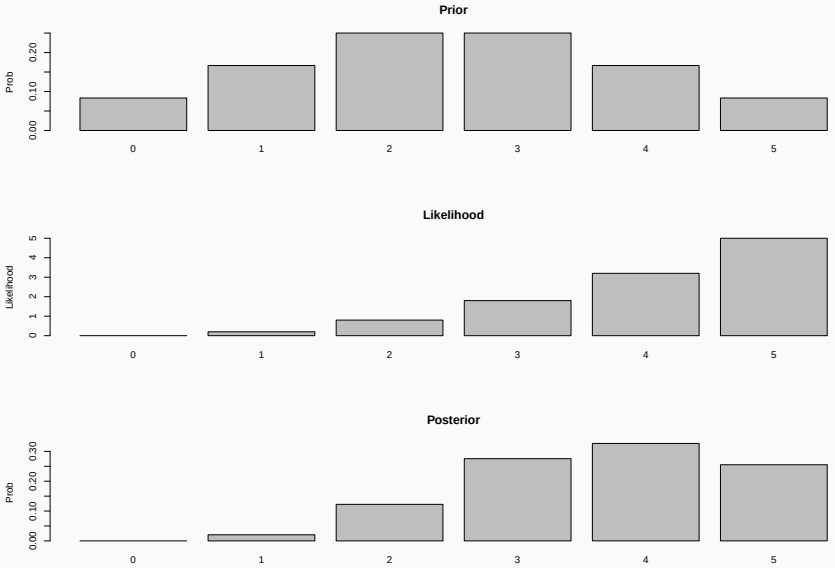


Fish Example: All data at once

You get the same result if the likelihood has both observations:

Num Red	Prior	Likelihood	Prior $\times$ Lik.	Posterior
0	1/12	0	0/300	0/98
1	2/12	(1/5)^2	2/300	2/98
2	3/12	(2/5)^2	12/300	12/98
3	3/12	(3/5)^2	27/300	27/98
4	2/12	(4/5)^2	32/300	32/98
5	1/12	1	25/300	25/98
Total	**1**	**3**	**98/300**	**1**

Fish Example: All data at once



Analytic Bayesian Methods: Conjugate Priors

Interrelation Among Distributions

Since the act of doing Bayesian analysis involves working with probability distributions, it's helpful to develop familiarity with individual distributions and relationships between distributions.

Uniform-Exponential-Gamma

- If $X \sim \text{Unif}(0, 1)$ then $-1/\lambda \log(X) \sim \text{Exp}(\lambda)$
- If $X_i \sim \text{Exp}(\lambda)$ the $\sum_i X_i \sim$ Gamma

Bernoulli-Binomial-Poisson-Beta

- If $X_i \sim$ Ber then $\sum_i X_i \sim$ Binomial
- As $n \to \infty$ then Binomial $\to$ Poisson
- Beta is a generalization of the Binomial

Cauchy-T-Normal-ChiSquared

- T with $\nu = 1$ is Cauchy
- T with $\nu \to \infty$ is Normal
- If $X \sim N(0, 1)$ then $X^2 \sim \chi^2_1$

. . . and many, many more!

Conjugate Priors

Before computers could help us determine the posterior, Bayesian analysis was conducted analytically (ie, with paper, pencil, and lots of algebra).

Given a particular form of the likelihood, there can be **convenient** choices for the form of the prior that make the math easier.

The easiest such group are called **Conjugate Priors**. This is when the prior and posterior are from the same family of distributions. Some examples:

- Beta-binomial
- Gamma-Poisson
- Normal-Normal

Conjugate Prior Example: Beta-Binomial

Suppose you work at a hedge fund whose trading model is based on major news announcements per week (X) from the n publicly traded firms in your portfolio. You want to estimate the probability of a major news announcement (θ).

The binomial likelihood is:

$$\mathbb{P}(X = k|n, \theta) = \binom{n}{x}\theta^k(1 - \theta)^{n-k} \quad \propto \quad \theta^k(1 - \theta)^{n-k}$$

Suppose we use a Beta(α, β) prior for θ:

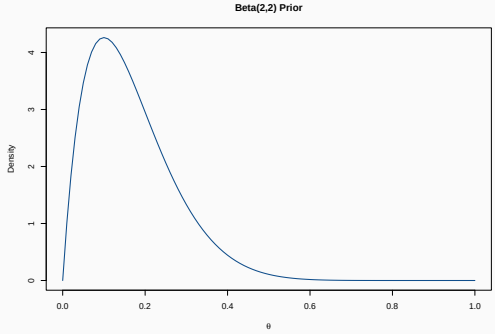
$$\mathbb{P}(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{B(\alpha, \beta)} \quad \propto \quad \theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

Then the posterior is the Beta($\alpha + k$, $\beta + n - k$) distribution!

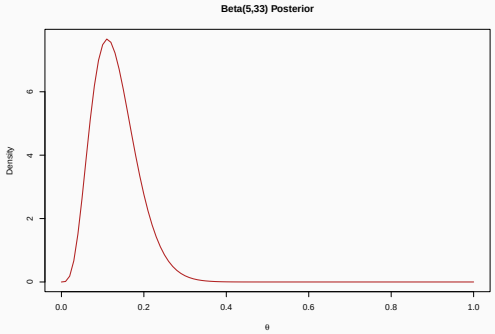
$$\begin{aligned}\mathbb{P}(\theta|\text{data}) &\propto \mathbb{P}(\text{data}|\theta) \times \mathbb{P}(\theta) \\ &\propto \theta^k(1 - \theta)^{n-k} \times \theta^{\alpha-1}(1 - \theta)^{\beta-1} \\ &\propto \theta^{(k+\alpha)-1}(1 - \theta)^{(n-k+\beta)-1}\end{aligned}$$

Conjugate Prior Example: Beta-Binomial

For the prior, suppose you specify a Beta(2,10) prior for θ to reflect your beliefs that major news announcements are relatively infrequent, but not rare:



For the likelihood, suppose you observe 3 major announcements out of 26 firms in your portfolio (11.5%). The posterior is then a Beta(2+3,26-3+10)=Beta(5,33) distribution:



Conjugate Prior Example: Gamma-Poisson

Suppose you work at a hedge fund whose trading model is based on the number of trades (Y) executed each second on a particular exchange. You want to estimate the rate (λ) of trades from n seconds of observed behavior.

The Poisson likelihood is:

$$\mathbb{P}(Y = y|\lambda) = \prod_{i=1}^{n} \lambda^{y} \exp\{-\lambda\}/y! \quad \propto \quad \lambda^{n\bar{y}} \exp\{-n\lambda\}$$

Suppose we use a gamma prior for λ:

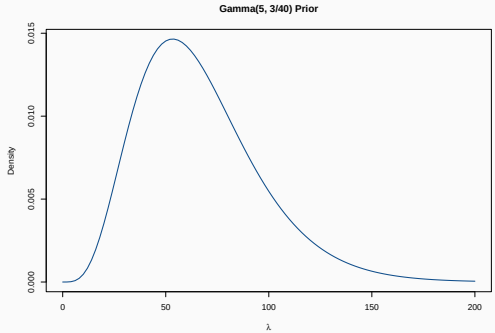
$$\mathbb{P}(\lambda|\alpha, \beta) = \beta^{\alpha}\lambda^{\alpha-1} \exp\{-\beta\lambda\}\Gamma(\alpha) \quad \propto \quad \lambda^{\alpha-1} \exp\{-\beta\lambda\}$$

Then the posterior is the Gamma($\alpha + n\bar{y}$, $\beta + n$) distribution!

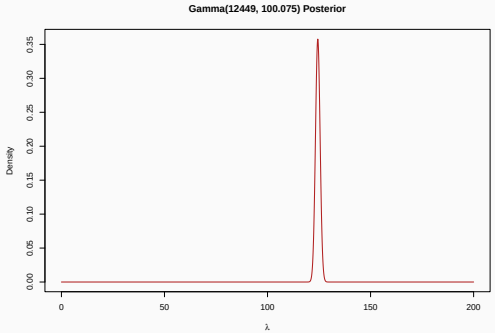
$$\begin{aligned}
\mathbb{P}(\lambda|y) &\propto \mathbb{P}(y|\lambda) \times \mathbb{P}(\lambda) \\
&\propto \lambda^{n\bar{y}} \exp\{-n\lambda\} \times \lambda^{\alpha-1} \exp\{-\beta\lambda\} \\
&\propto \lambda^{\alpha+n\bar{y}-1} \exp\{-(\beta + n)\lambda\}
\end{aligned}$$

Conjugate Prior Example: Gamma-Poisson

For the prior, suppose you specify a Gamma(5, 0.075) prior for λ to reflect your beliefs that trading on the focal exchange is similar to trading frequency on the NYSE (with 1.5mil trades/day or 52 trades/sec):



For the likelihood, suppose you observe 100 seconds of trading activity capturing 12,444 trades (124.44 trades/sec). The posterior is then a Beta(5+12444,3/40+100)= Beta(12449,100.075) distribution:



Posterior Sampling Methods: MCMC

Leaving conjugates behind

Suppose you have a high-dimensional (many parameter) problem, or simply one where your prior is not a conjugate to your likelihood.

How can you calculate or approximate the posterior?

- Grid approximation*
- Quadratic approximation
- MCMC: Markov Chain Monte Carlo (Metropolis*, Gibbs, Hamilton)

Grid Approximation

Use the ideas of the Bayes' Tables on slides 8 and 10.

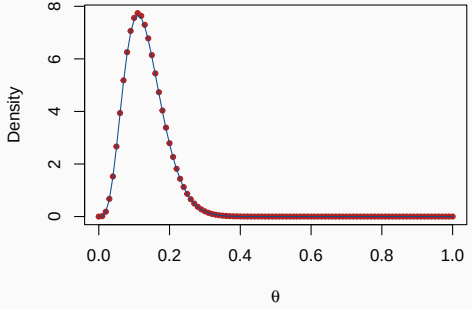
- Take a sequence of values in the parameter space
- For each value, calculate the prior and likelihood
- Calculate the denominator as the mean of the prior-times-likelihood values
- Scale the prior-times-likelihood values by the denominator to get the posterior

Note:

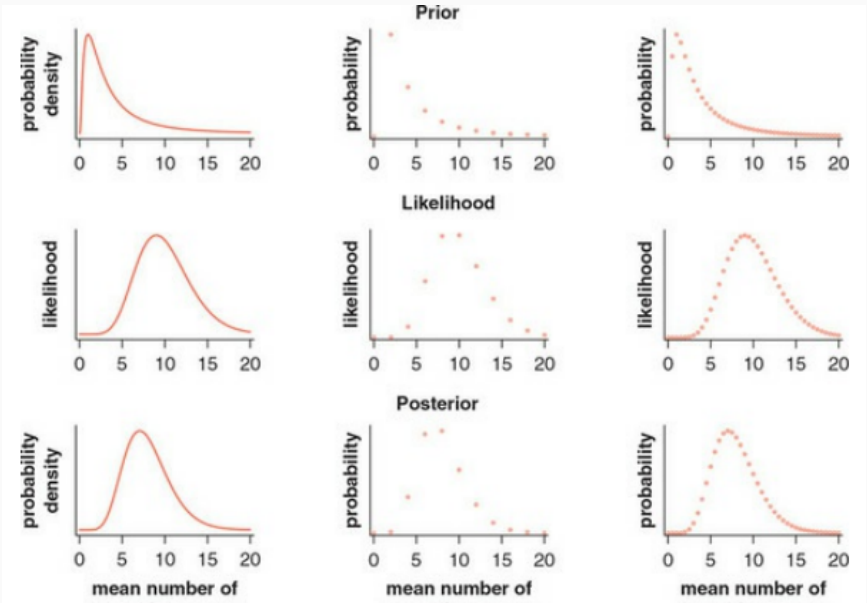
- For discrete distributions, this *is* how to calculate the posterior
- For continuous distributions, this offers a good approximation
- Note that this becomes computationally intractable for high dimensional problems

Grid Approximation for Beta-Binomial Example

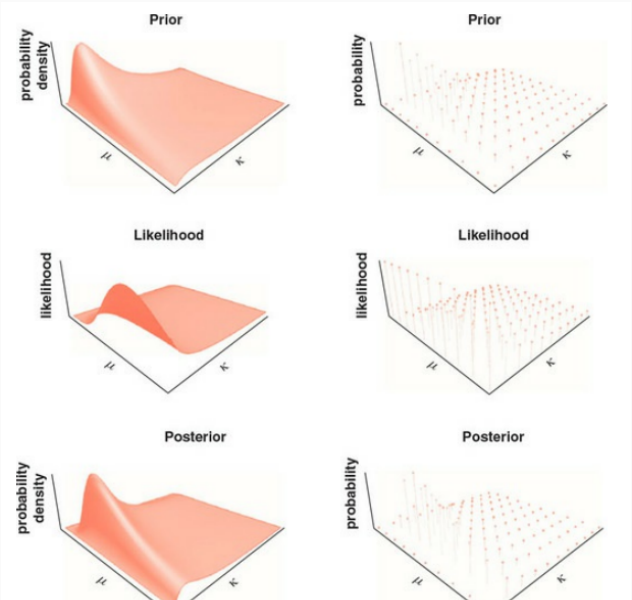
```r
# grid approx calcs
s <- seq(0, 1, by=0.01)
prior <- dbeta(s, 2, 10)
lik   <- dbinom(x=3, size=26, prob=s)
denom <- mean(prior*lik)
post  <- lik * prior / denom

# grid approx as red points
plot(s, post, pch=20, col="firebrick",
     ylab="Density", xlab=expression(theta))

# true posterior as blue line
lines(s, dbeta(s, 5, 33), col="dodgerblue4")
```



Grid Approximation 1D example



Grid Approximation 2D example



MCMC: Markov Chain Monte Carlo

A **Markov Chain**is a "memoryless" stochastic process.

The Metropolis-Hastings algorithm is a general method to simulate a target distribution (often denoted π) by defining a Markov Chain with stationary distribution π.

This means that, in the long run (ie, over repeated or Monte Carlo simulations), samples from the Markov Chain look like samples from π.

Popular special cases of MH MCMC include:

- Metropolis*
- Gibbs*
- Hamiltonian Monte Carlo (HMC)

Metropolis

The Metropolis algorithm is quite simple for a single parameter:

- Initialize θ at a specific value, say θ_0
- For $t = 1, 2, \ldots$
  - Sample θ^* from some symmetric distribution Q_t
  - Calculate the acceptance probability $A = \pi(\theta^*)/\pi(\theta_t)$
  - Move (ie, set $\theta_{t+1} = \theta^*$) with probability A
  - Alternatively, stay (ie, set $\theta_{t+1} = \theta_t$) with probability 1 - A

Animated example here:

https://youtu.be/Qqz5AJjyugM?t=360

Gibbs Sampling

For multivariate distributions where $\pi(\theta^{(2)}|\theta^{(1)})$ is easy to compute:

- Initialize $\theta = (\theta_0^{(1)}, \theta_0^{(2)})$
- For $t = 1, 2, \ldots$
  - Then draw $\theta_t^{(1)}|\theta_{t\text{-}1}^{(2)}$
  - Then draw $\theta_t^{(2)}|\theta_t^{(1)}$
  - Repeat the conditional draws until $t = T$

Demonstrated in R with `bayesm::rbiNormGibbs(rho=0.5)`

More Bayes

Truly excellent online video lecture series:

- Grant Sanderson's 3B1B Bayes Video [link]
- Richard McElreath's Statistical Rethinking Lectures [link]
- Ben Lambert's Bayesian Statistics Lectures [link]

Bayesian Stats textbooks:

- *Bayesian Data Anlalysis* by Gelman et al.
- *Statistical Rethinking* by McElreath
- *Doing Bayesian Data Analysis* by Kruschke
- *Bayes Rules!* by Johnson, Ott, and Dogucu
- *A Student's Guide to Bayesian Statistics* by Lambert
- *Bayesian Essentials with R* by Marin and Robert
- *The Bayesian Choice* by Robert

Course Wrap Up

Final Exam

- Tuesday, December 10th from 11:30am to 2:30pm
- You may bring:
  - Two 8.5" x 11" sheets of paper with notes on both sides
  - A calculator (cannot connect to the internet)
  - Your ID
- My intentions are to make an exam that is:
  - Cumulative and comprehensive
  - Slightly longer than the midterm
  - Most of all: fair, emphasizing core concepts and techniques

Course Evaluation

Please take a few minutes to give detailed feedback about the course.

- What did you like?
- What can be improved?

Please note that:

- The course evaluations are anonymous
- I will not see the results until after I submit final grades
- I will read every comment and take them seriously
- Your feedback will be used to improve the course in the future

All of my past course evaluations for MFE 402 and all other courses I've taught are available on my website:

- https://www.danyavorsky.com/teaching

Penultimate Slide

Give this course one last big effort for the final exam... then:

Be proud! Congratulate yourself! You have learned a lot of material in a short amount of time.

You have:

- Simulated the WLLN and the CLT
- Multiplied a lot of matrices
- Computed heteroskedastic standard errors
- Plotted 100 confidence intervals
- Thought about leverage, omitted variable bias, and multicollinearity
- Used linear regression techniques for causal inference
- Coded and maximized a log-likelihood function
- Bootstrapped standard errors
- Simulated a Bayesian posterior distribution

What Comes Next

- **Time Series Analysis:** when your data are not independent
- **Machine Learning:** when prediction overshadows inference
- **Nonparametrics:** when you have lots of data and make fewer assumptions about $m(X)$
- **Instrumental Variables:** a tool for causal inference and/or to combat endogeneity
- **Extending the topics we covered:**
  - Other generalized linear models (Tobit for truncated data)
  - Other Bayesian models (hierarchical models)
  - Mixing time series and cross-sectional (panel data)

Please Keep in Touch

I want to know:

- What's a great class you've taken?
- What did you read that is super interesting?
- Are you working on interesting questions in your internship?
- Are you using interesting methods at your job?
- What is it like to work at your company?

Contact Info:

- dyavorsky@gmail.com
- linkedin.com/in/dyavorsky
- 951-201-0927