Introduction to Bayesian Statistics

Statistical Learning Workshop

Kevin McAlister

University of Michigan

November 13, 2017

A Motivating Example

- Sam writes down a number and flips a coin.
- If the coin turns up heads, then Sam tells Jen if the number is even or odd.
- If the coin turns up tails, Sam lies to Jen.
- Jen then guesses if Sam's number is even or odd.
- Let θ be the probability that Jen correctly guesses if Sam's number is even or odd.

A Motivating Example

- Before any data has been collected:
  1. What is our best guess about θ?
  2. What is the probability that $\theta > \frac{1}{2}$?
- After data has been collected:
  1. What is our best guess about θ?
  2. What is the probability that $\theta > \frac{1}{2}$?

Statistics as a Whole

- Statistics is largely the study of quantifying **uncertainty**.
- Models are used to describe and predict outcomes from data.
- Explore meaningful hypotheses using observable data.
- Much of the burden of data analysis arises in describing the quality of inferences made from data.

The Frequentist Paradigm

- A **frequentist** approach quantifies uncertainty in terms of repeating the procedure that generates the data a large number of times.
- Parameters of interest (θ) are fixed and unknown. These are what we are inferring using models.
- The sample data (Y) is random.
- A frequentist approach **never** views θ as a random variable.
- $P(\theta > x) =$???
- All probability statements are made about randomness in the data.

The Frequentist Paradigm

- Inference is made using a statistic ($\hat{\theta}$), which is a function of the data.
- This statistic should be representative of θ (Consistent, Efficient, Sufficient, etc.)
- $\hat{\theta}$ has a **sampling distribution** - the distribution of uncertainty associated with $\hat{\theta}$ due to randomness in the data.
- θ does not have a distribution.

The Frequentist Paradigm

- A common approach for testing hypotheses is to reject the null if a test statistic exceeds some threshold.
- For example, for $H_0 : \theta = 0$ reject H_0 if $|\hat{\theta}| > \alpha$.
- A **p-value** is the probability of observing a test statistic as extreme as observed if sampling repeated a large number of times.
- Inverting the above test yields a $\alpha\%$ confidence interval.
- A confidence interval should contain the true value of θ $\alpha\%$ of times we repeat sampling.

The Frequentist Paradigm

- Frequentist tests never say the probability that the null is true.
- Frequentist confidence intervals have no probabilistic interpretation - the probability that a single confidence interval contains the true value of θ is 0 or 1.

Back to the Motivating Example

- From a frequentist perspective.
- Before any data has been collected:
  1. What is our best guess about θ? *No Idea*
  2. What is the probability that $\theta > .5$? *θ isn't a random variable, so it doesn't have a distribution.*
- After data has been collected:
  1. What is our best guess about θ? *The sample proportion.*
  2. What is the probability that $\theta > .5$? *This is a nonsense question. θ isn't a random variable.*

Why does any of this matter?

"Then you should say what you mean," the March Hare went on.

"I do," Alice hastily replied; "at least–at least I mean what I say–that's the same thing, you know."

"Not the same thing a bit!" said the Hatter. "You might just as well say that "I see what I eat" is the same thing as "I eat what I see"!

— Lewis Carroll

Why does any of this matter?

- There's a time and a place for everything – frequentist statistics aren't inherently bad.
- Speaking frequentism is often not what we want to say.
- With models, we often want to make *probabilistic* claims (i.e. the probability that $\theta > 0$ is greater than .95).
- Assumptions are often obscured and difficult to justify.
- There is often **prior** knowledge that we want to introduce into analysis.
- Not all questions fit in the strict statistical hypothesis framework.

Why does any of this matter (especially to a social scientist)?

- Where does frequentism fit when the idea of resampling is impossible?
- Genocide, elections, market crashes, etc. are all isolated events that cannot be recreated by the scientific process.
- What about Congressional roll call votes? We have the entire accurate sample. How can uncertainty be related to sampling?
- These are important philosophical concerns that relate to how we interpret results.
- Rigid definition of probability leads to poor inference and, in turn, poor science.

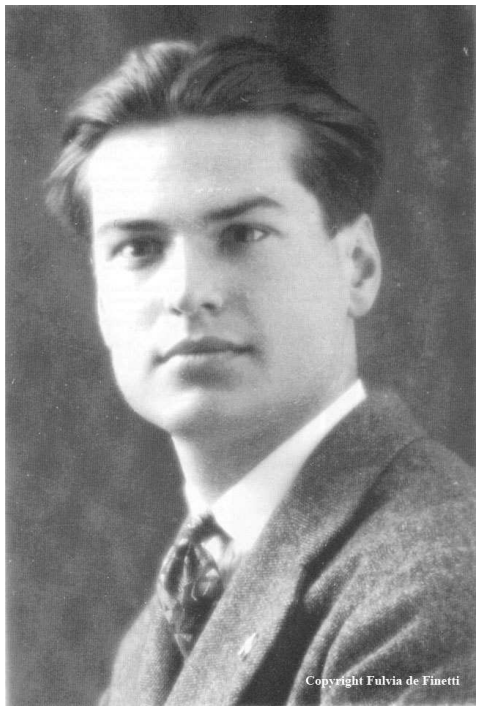
De Finetti Drops the Mic

PROBABILITY DOES NOT EXIST: The abandonment of superstitious beliefs about...Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is not less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs. In investigating the reasonableness of our own modes of thought and behaviour under uncertainty, all we require, and all that we are reasonably entitled to, is consistency among these beliefs, and their reasonable relation to any kind of relevant objective data (relevant in as much as subjectively deemed to be so). This is Probability Theory.

— Bruno de Finetti

Subjective Uncertainty

- Sometimes, probability should reflect the strength of belief that something is true.
- This is in stark contrast to objectivist notions of probability – probability is no longer assumed a property of the object under study.
- Bayesian methods provide structured rules for quantifying *subjective* uncertainty.
- Bayes theorem provides a method for updating our beliefs about θ after observing data.



Beliefs as Distributions

- θ is a fixed but unknown feature of the population from which data is being samples.
- There is a true θ and sampled data is a function of this true value.
- However, θ is only observable as a random variable with subjective uncertainty.
- Our beliefs about the value of θ are conditional on data and can be represented by probability, $P(\theta|Y)$.

Beliefs as Distributions

- Beliefs must follow the rules of probability distributions:
  1. $\int_{-\infty}^{\infty} P(\theta|Y)d\theta = 1$
  2. $P(\theta|Y) \geq 0 \; \forall \; \theta \in (-\infty, \infty)$
- $P(\theta|Y)$ is known as the **posterior** distribution of θ. The posterior is the object of interest for Bayesian inference.

Posteriors and Subjective Uncertainty

- Posterior distributions are a full characterization of our subjective uncertainty about θ after looking at data.
- It contains everything we need for making inferences about the parameters.
- Examples:
  - The posterior probability that a regression coefficient is positive, negative, or lies in a particular interval
  - The posterior probability that a subject belongs to a particular latent class
  - The posterior probability that a hypothesis is true
  - The posterior probabilities that a particular statistical model is true model among a family of statistical models

Posteriors and Subjective Uncertainty

- Note that all of these examples are questions that cannot be directly answered under the frequentist paradigm.
- Allowing uncertainty to exist with the parameter allows for a more natural interpretation of inference about the parameters.
- Interpretation is intuitive - we can make the probabilistic claims that we want with confidence intervals and p-values.
- Repeated sampling is not necessary for this interpretation. Everything can be interpreted with respect to the observed sample.
- There is a cost! We have to be very explicit about what assumptions we're making. This is both a good and bad thing.

Conditional Probability

- Let A and B be events where $P(A) > 0$ and $P(B) > 0$.
- The conditional probability of A given B occurs is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Multiplication Rule:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- Law of Total Probability:

$$P(B) = P(B|A)P(A) + P(B| \sim A)P(\sim A)$$

Bayes theorem

- From these rules of probability, we can derive Bayes Law:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Assume A and B follow probability distributions. In the discrete case:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{A} P(B|A)P(A)}$$

- In the continuous case:

$$P(A|B) = \frac{P(B|A)P(A)}{\int_{-\infty}^{\infty} P(B|A)P(A)dA}$$

Bayes theorem for Data

- Bayes theorem is always true and can be applied to arbitrarily complex situations.
- We want to estimate the posterior distribution of θ given Y, $P(\theta|Y)$.
- Applying Bayes theorem:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} = \frac{P(Y|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(Y|\theta)P(\theta)d\theta}$$

- In words:

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Likelihood

- The part of this formula that integrates data is the **likelihood**.
- Make an assumption about the data generating process of the observed data (i.e. $P(Y|\theta) \sim PDF(.)$).
- Define Y as a sample of size N from the population of interest, $y_i \ \forall \ i \in (1, N)$
- If we assume y_i is an i.i.d sample, then

$$P(Y|\theta) = \prod_{i=1}^{N} P(y_i|\theta)$$

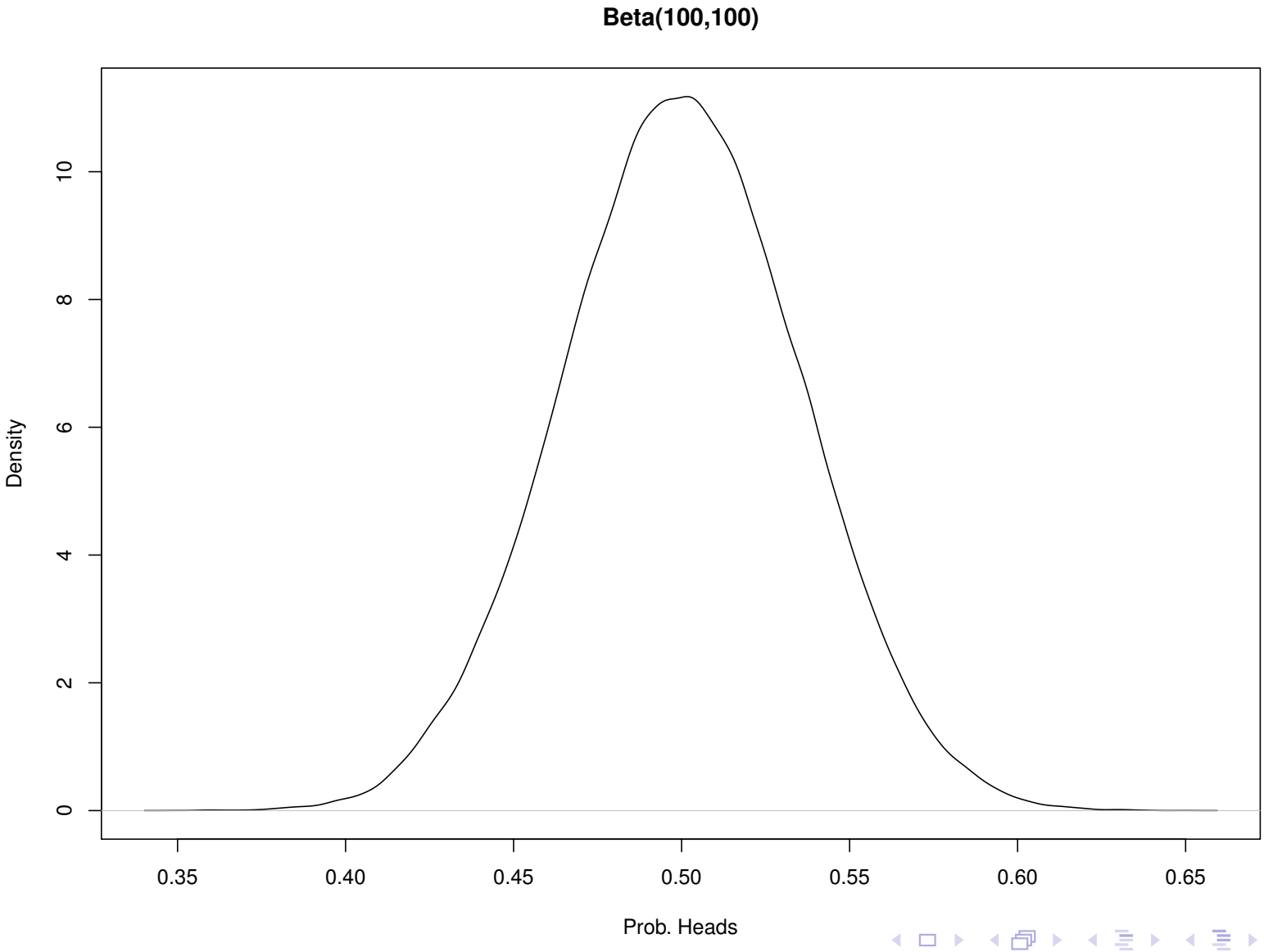
- Bayesian methods are very similar to likelihood estimation. This will be seen more explicitly later.

Priors

- The second part of the Bayesian formula.
- $P(\theta)$ is a distribution that quantifies out *a priori* beliefs about θ.
- Let θ be the probability that I flip a coin and it turns up heads. I believe there's a strong chance the coin is fair and $\theta = .5$. I can define my prior distribution on θ as:

$$P(\theta) \sim Beta(100, 100)$$

Priors


Beta(100,100)
Density
0
2
4
6
8
10
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Prob. Heads

Priors

- The selected prior should quantify your prior beliefs about the value of θ.
- Sometimes, this can be elicited via previous research or experts.
- Most of the time, however, this is selected to be mathematically convenient and/or diffuse.
- We'll discuss this more later.

Marginal Likelihood

- The final part of the formula is the model evidence or marginal likelihood, $P(Y)$.
- Marginalize out the prior to know how well the data is described by the model.
- Used for nested and non-nested model comparison.
- Often very difficult to calculate analytically.
- $P(Y)$ is a *normalizing constant* that scales the posterior distribution.
- While the numerator is always a function of θ, the denominator is never a function of θ.

Back to the Motivating Example (Finally)

- We observe N different groups play the coin game. We recorded the N outcomes - $y_i \in (0, 1)$.
- Each outcome is and independent and identically distributed draw from a Bernoulli distribution:

$$P(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

- We want to find the posterior distribution of θ given N samples.
- θ must be between 0 and 1. Let $P(\theta) \sim Unif(0, 1)$.
- $P(\theta) = 1$
- Let's find the posterior distribution of θ.

Example Likelihood

- Our N samples are i.i.d draws from a Bernoulli distribution.
- Likelihood:

$$P(Y|\theta) = \prod_{i=1}^{N} P(y_i|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum y_i}(1-\theta)^{N-\sum y_i}$$

Example Posterior

- Using Bayes theorem, the posterior is:

$$P(\theta|Y) = \frac{\theta^{\sum y_i}(1-\theta)^{N-\sum y_i}}{\int_0^1 \theta^{\sum y_i}(1-\theta)^{N-\sum y_i} d\theta}$$

- This integral is hard to solve. Fortunately, this is a well known integral with a known solution:

$$\int_0^1 \theta^{\sum y_i}(1-\theta)^{N-\sum y_i} d\theta = \frac{\Gamma(\sum y_i + 1)\Gamma(N - \sum y_i + 1)}{\Gamma(N+2)}$$

which is the Beta function - $\mathbf{B}\left(\sum y_i + 1, N - \sum y_i + 1\right)$

Example Posterior

- This is a known distribution!

$$P(\theta|Y) \sim Beta\left(\sum y_i + 1, N - \sum y_i + 1\right)$$

- This isn't a convenient coincidence.
- This is due to **conjugacy** of the prior.

Conjugate Priors

- Let the likelihood have a known distribution - $P(Y|\theta) \sim f(Y;\theta)$
- A conjugate prior is a distribution, $P(\theta) \sim g(\theta;.)$, s.t.:

$$g(\theta; Y, .) = \frac{f(Y;\theta)g(\theta;.)}{\int_{\infty}^{\infty} f(Y;\theta)g(\theta;.)d\theta}$$

- In words, the posterior has the same distributional form as the prior.
- This is really important! Guarantees that the posterior has a known form. Math becomes much easier.
- All distributions in the exponential family have a conjugate prior.

The Bayesian Mantra

- In the example, we saw that θ was only involved in the numerator of the posterior.
- The denominator is always a constant.
- Using some algebra:

$$\int_{-\infty}^{\infty} P(Y|\theta)P(\theta)d\theta = P(Y)$$

- The numerator is simply an unscaled probability distribution.
- We really don't need to calculate the marginal likelihood to characterize the posterior.
- This leads to the "Bayesian mantra":

$$Posterior \propto Likelihood \times Prior$$

What does the prior really do?

- This statement of the posterior makes the role of the prior very clear.
- The posterior is a compromise between the information learned from the data and our prior beliefs.
- The posterior can be sensitive to prior choice:
  1. If the location of the prior and the likelihood agree, then there is little influence.
  2. If the location of the prior and the likelihood are different, then the prior is influential.
  3. If the variance of the prior is much lower than the variance of the likelihood, then the prior choice dominates the posterior.

Coin Example Revisited

- Let's revisit the coin example.
- $P(Y|\theta) = \theta^{\sum y_i}(1-\theta)^{N-\sum y_i}$
- The conjugate prior for the binomial distribution is the beta distribution - $P(\theta) \sim Beta(\alpha, \beta)$.

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

- Let's find the posterior for θ given an arbitrary beta prior

Coin Example Revisited

- The prior is conjugate, so we know that the posterior is beta distributed.

$$P(\theta|Y) \propto \frac{1}{\mathbf{B}(\alpha,\beta)}\theta^{\sum y_i}(1-\theta)^{N-\sum y_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

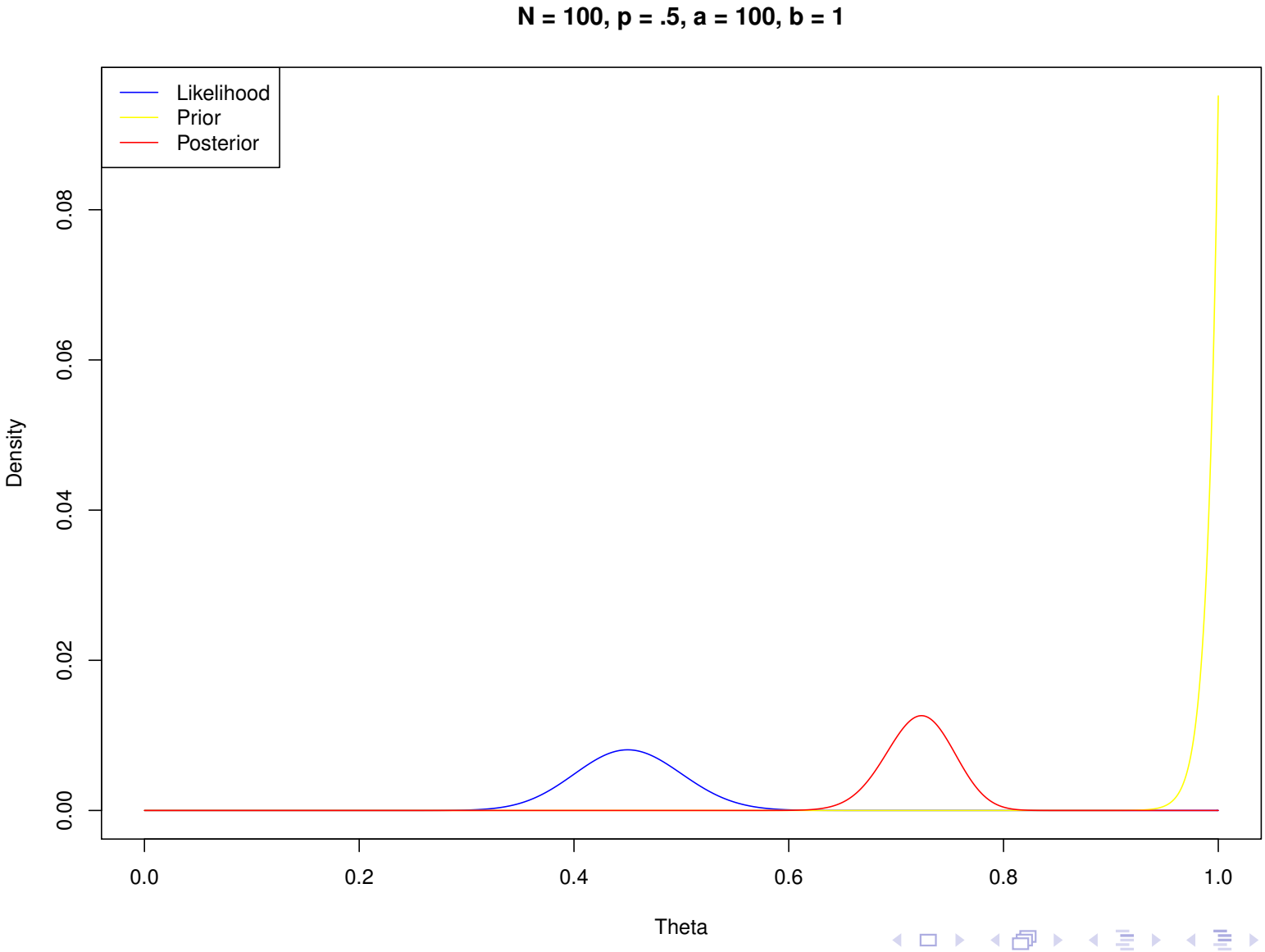
$$P(\theta|Y) \propto \theta^{\sum y_i}(1-\theta)^{N-\sum y_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$P(\theta|Y) \propto \theta^{\sum y_i+\alpha-1}(1-\theta)^{N+\beta-\sum y_i-1}$$

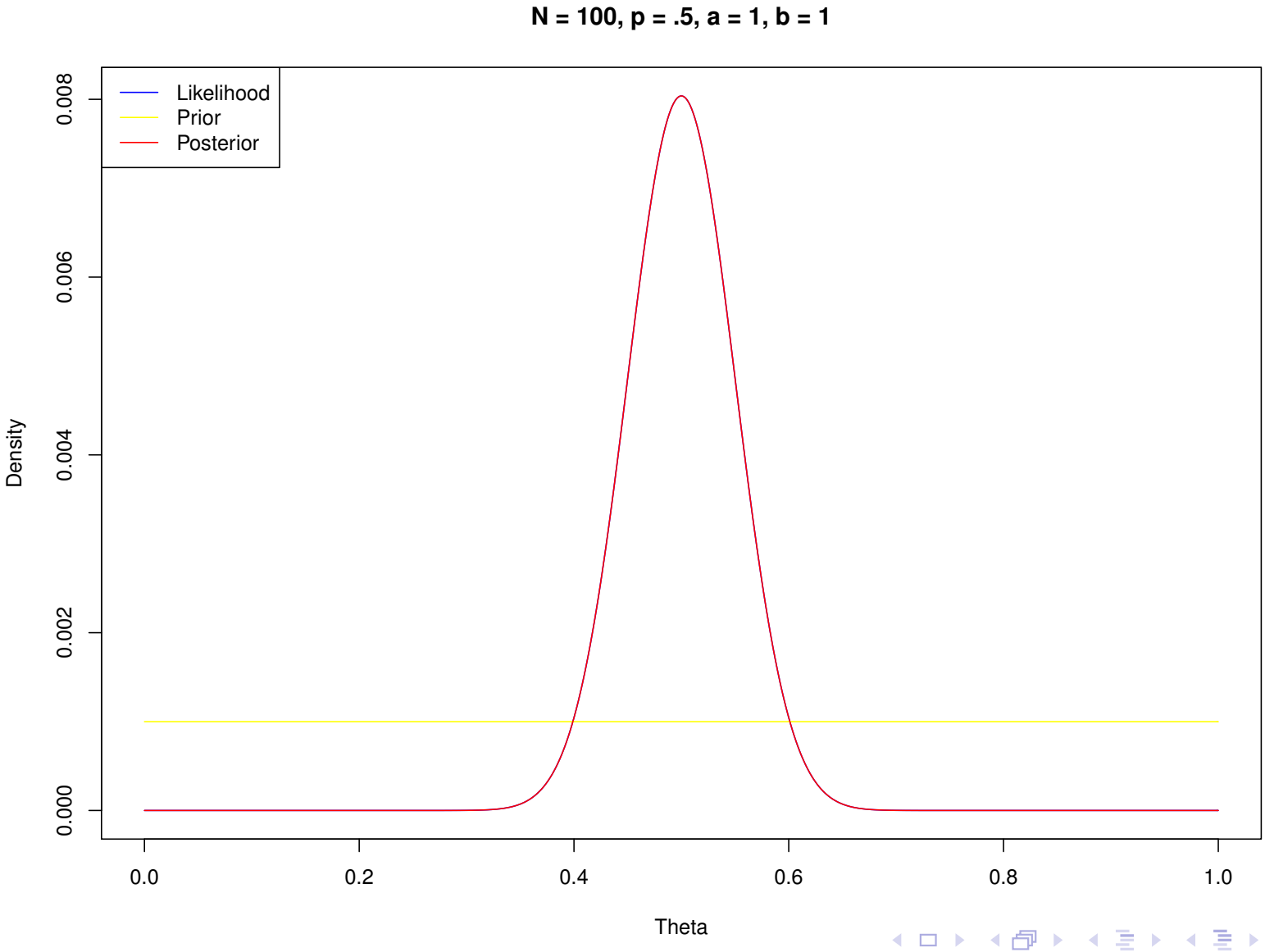
- The posterior is

$$P(\theta|Y) \sim Beta\left(\sum y_i + \alpha, N - \sum y_i + \beta\right)$$

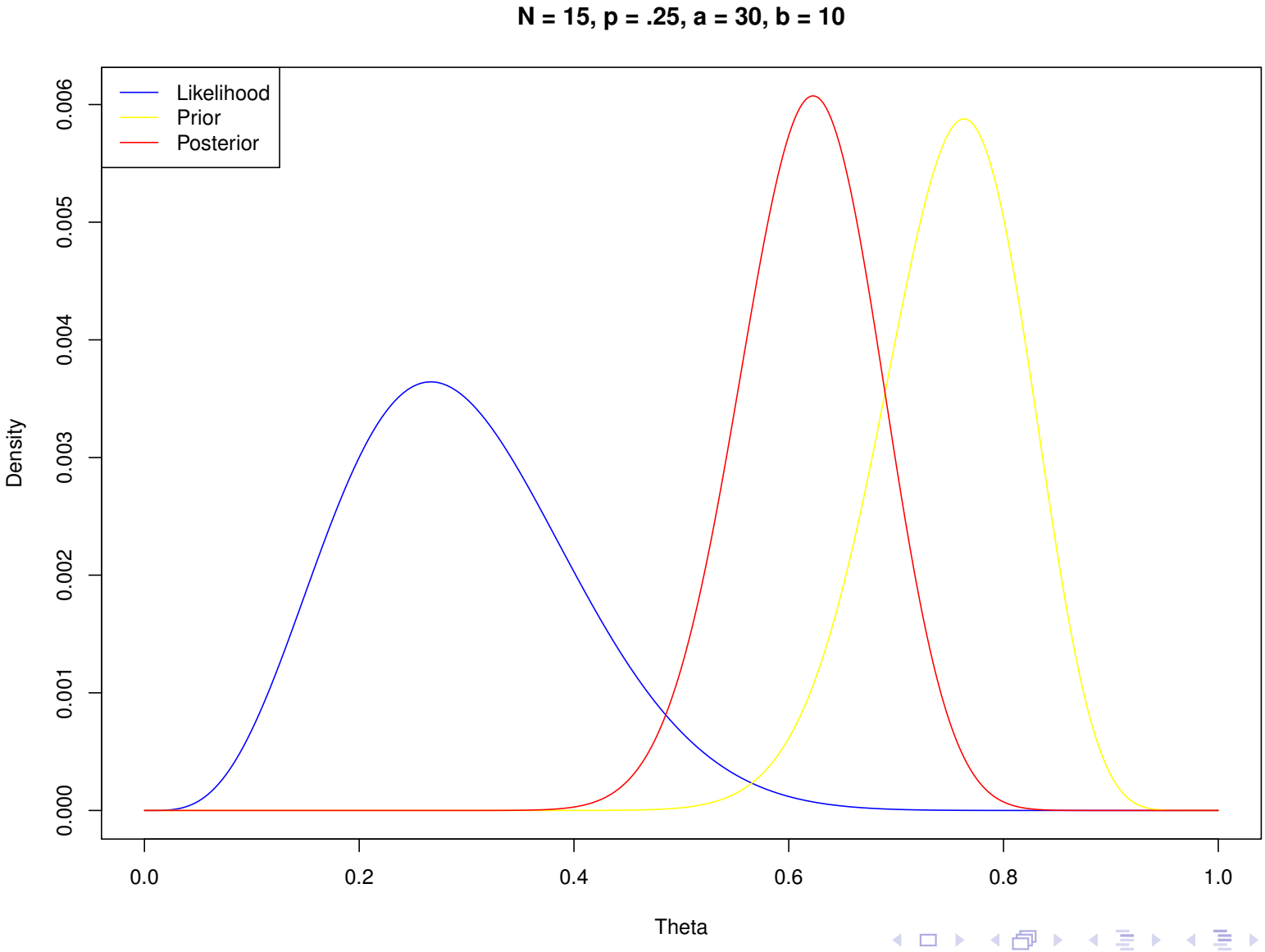
Coin Example Revisited



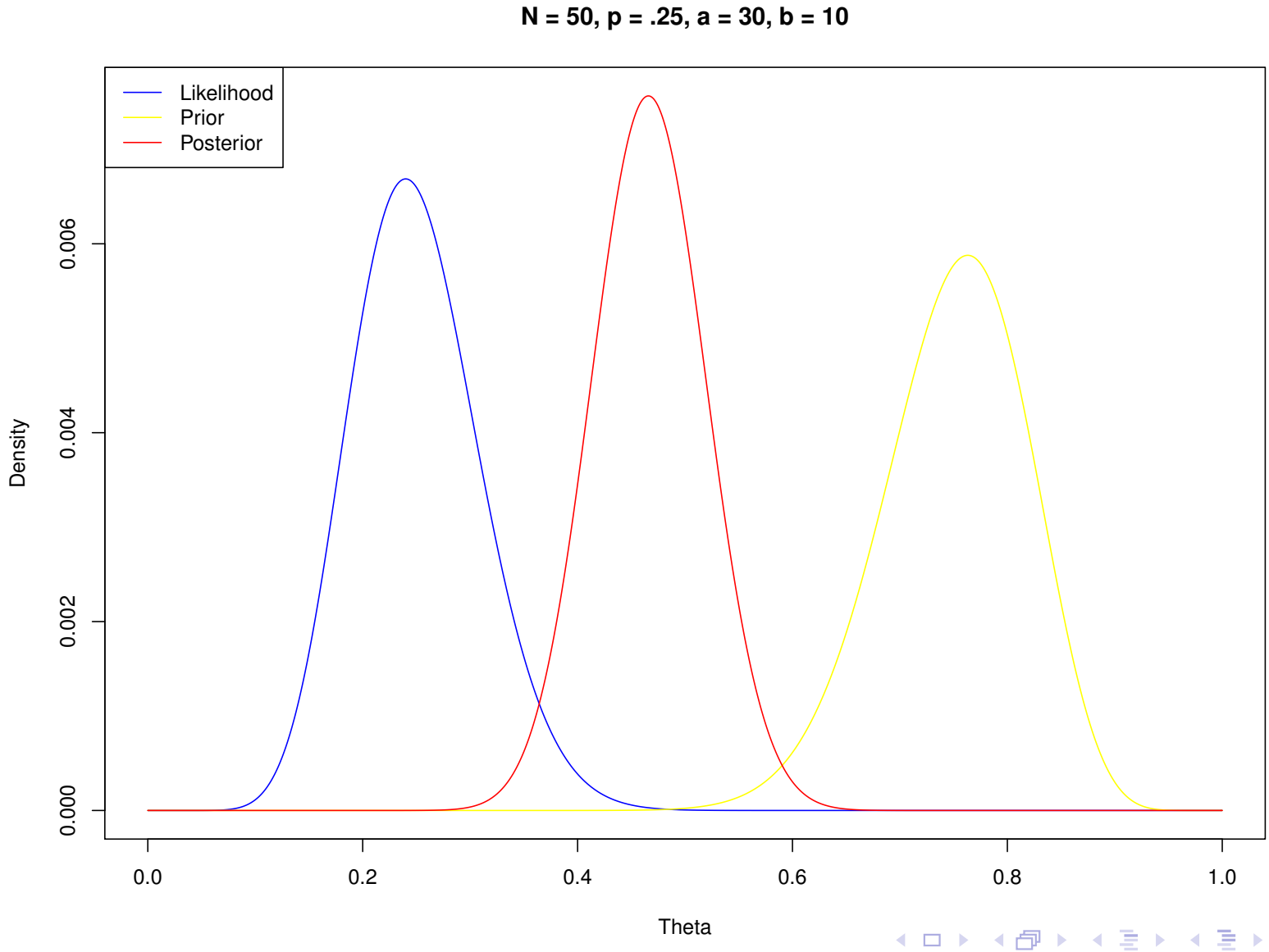
Coin Example Revisited



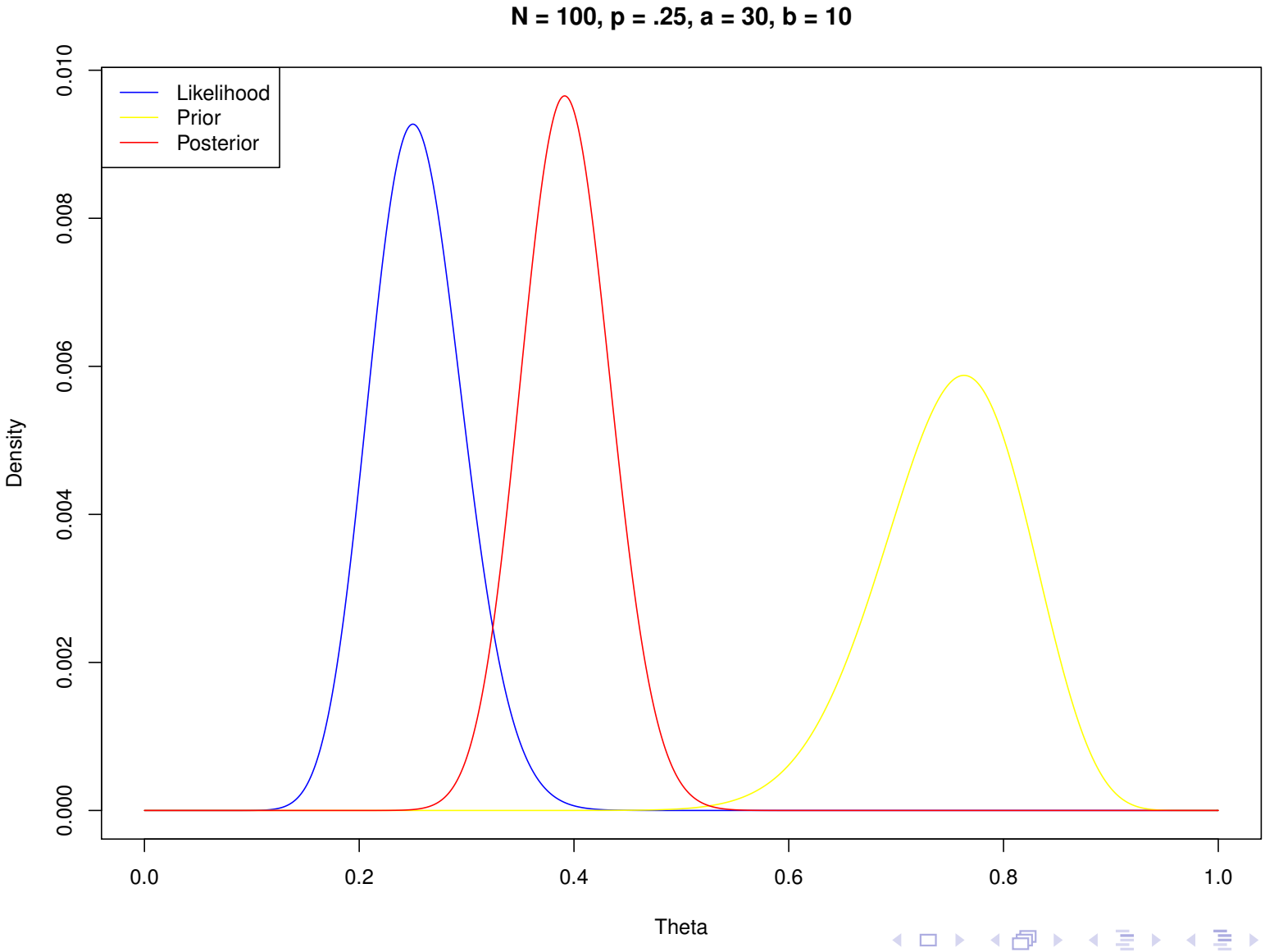
Coin Example Revisited



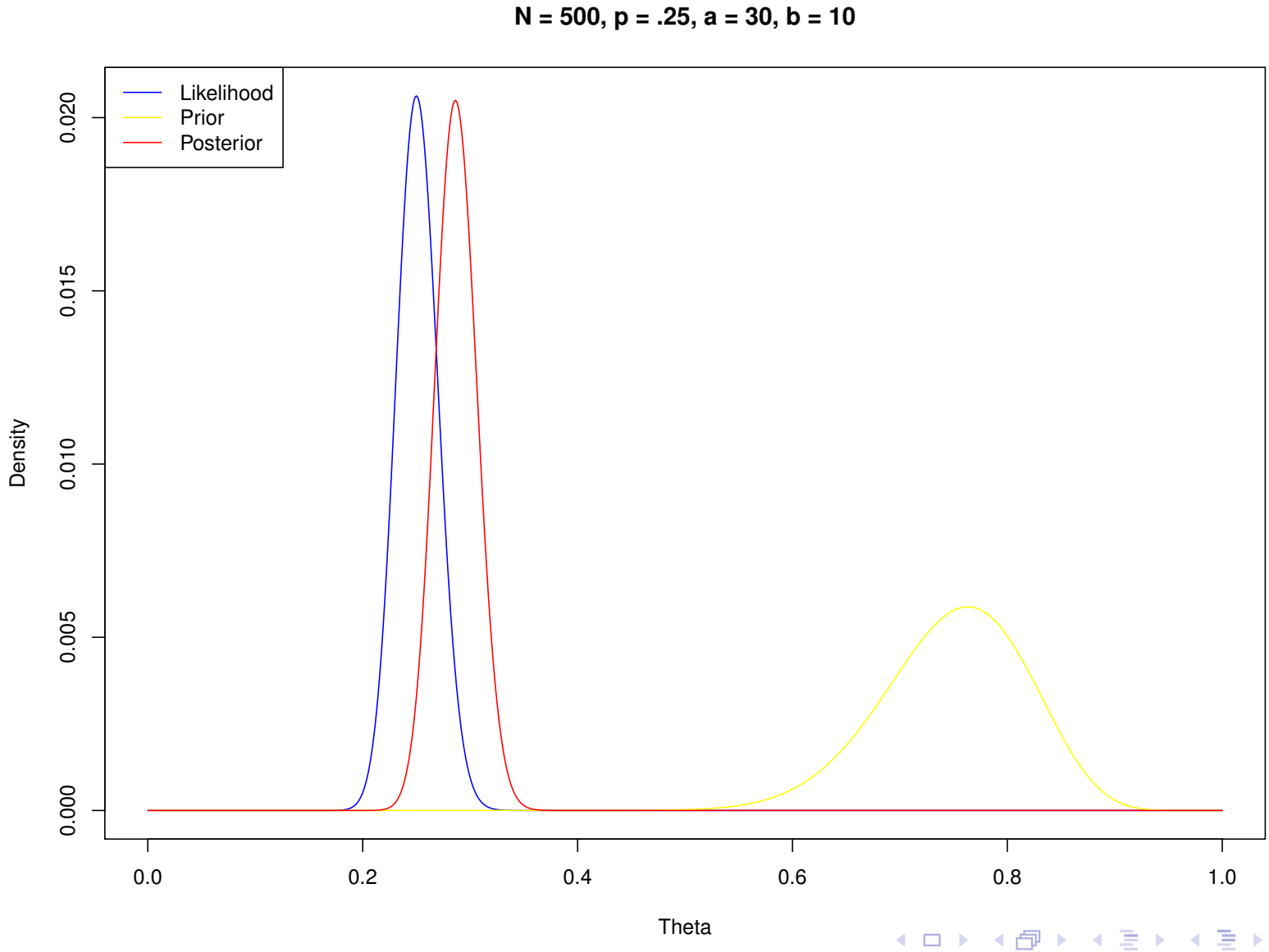
Coin Example Revisited



Coin Example Revisited



Coin Example Revisited



Consistency and Frequentist Duality

- For most posterior distributions, as $N \to \infty$, the likelihood completely dominates the prior and prior specification doesn't matter.
- As $N \to \infty$, $P(\theta|Y) \xrightarrow{D} N(\theta, f(1/N))$.
- The mean of the posterior distribution converges to the maximum likelihood estimate of θ as N gets large, regardless of prior distribution.
- This speaks to a duality between frequentist and Bayesian estimates in larger samples.
- A larger notion that there are commonalities between the two approaches.
- Use whichever approach makes the most sense when N is large.

Point Estimators and Bayes Loss

- Sometimes point estimators are useful.
- Minimize Bayes loss using loss functions.
- Quadratic Loss $\rightarrow$ Posterior Mean
- Absolute Loss $\rightarrow$ Posterior Median
- All-or-Nothing Loss $\rightarrow$ Posterior Mode
- Point estimators at the mode are very interesting and specific to the Bayesian approach - if I had to make a best guess, it would be the highest density point.

Credible Intervals

- The big advantage of Bayesian inference is the ability to make probabilistic claims about θ.
- Define a α% credible interval $[l, u]$ as:

$$\int_{l}^{u} P(\theta|Y)d\theta = \alpha/100$$

- We can interpret this interval as we want to interpret confidence intervals - the probability that θ falls in this interval is α.
- A reasonable interval to report is the smallest interval that covers α% of the posterior density. This is called the highest posterior density interval.
- There is no rule that says that this interval has to be continuous. This method allows meaningful inference for multimodal outcomes.

Bayesian Model Comparison

- Another critical distinction between frequentist and Bayesian approaches is the hypothesis space.
- Frequentists assume that the hypothesis space is infinite and test against a known, "uninteresting" hypothesis. Rejecting the null lets us know that the parameter of interest is "interesting".
- Bayesians assume that the hypothesis space is finite. By the normalization result, Bayesians define the hypothesis space as a discrete number of possibilities and calculate the probability that each hypothesis is correct.
- Assume that each model tests a unique hypothesis. We can compare models and select the best one given the data.
- The marginal likelihood, $P(X)$, is used as the assessment of the quality of a model. As $P(X)$ increases, the model does a better job of fitting the data.
- The key here is that $P(X)$ marginalizes the estimated parameters over the priors.

Bayesian Model Comparison

- Model comparison utilizes a Bayes Factor. When comparing two models (M_1 and M_2), the Bayes Factor is:

$$BF_{1:2} = \frac{P(X|M_1)}{P(X|M_2)}$$

- A Bayes Factor greater than 1 favors M_1. A bigger value has a higher probability of being correct.
- Given that a Bayes Factor expresses the odds that M_1 is better than M_2, we can extend the Bayes Factor to probabilities.
- Similarly, we can do multiple pairwise comparisons and compute probabilities for more than 2 models.

Bayesian Advantages

- Bayesian concepts (posterior prob of the null) are arguably easier to interpret than frequentist ideas (p-value)
- We can incorporate scientific knowledge via the prior
- Excellent at quantifying uncertainty in complex problems
- In some cases the computing is easier
- Provides a framework to incorporate data/information from multiple sources

Bayesian Disadvantages

- Picking a prior is subjective
- Procedures with frequentist properties are desirable
- Computing can be slow or unstable for hard problems
- Less common/familiar

The Evangelizing Slide

- Bayesian inference provides a lot of niceties for statistical analysis:
  - Provides probabilities for hypotheses
  - Simple interpretation
  - Explicit assumptions
  - Marginalizes nuisance parameters
  - Model comparisons for more than 2 nested or non-nested models
  - Automatic overfitting penalties via Occam's factors
  - Valid for all sample sizes
  - Handles multimodality
  - Accounts for prior information and tests
  - Does not suffer from early stopping of experiments
  - Provides consistent, calibrated estimators
  - Good coverage (frequently better than frequentist analogues)

Some Applications

- Let's move on to some problems.
- Like we've previously discussed, we want to use conjugate priors when at all possible.
- In survival analysis, the exponential distribution is frequently used to model how long an object lives.

Some Applications

- Let's assume that we have N observations that follow an exponential distribution:

$$P(x_i|\lambda) = \lambda \exp[-\lambda x_i]$$

where

$$\lambda > 0 \ \& \ x_i > 0$$

- We want to infer about the value of λ.
- The conjugate prior for the exponential distribution is the Gamma distribution:

$$P(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp[-\beta\lambda]$$

- What form will the posterior of λ take?

Some Applications

- Recall that we're after:

$$P(\lambda|x) \propto P(x|\lambda)P(\lambda)$$

- The likelihood is pretty simple:

$$P(x|\lambda) = \prod P(x_i|\lambda) = \lambda^N \exp\left[-\lambda \sum x_i\right]$$

- We can combine the likelihood and the prior by matching *kernels*.

$$P(\lambda|x) \propto \lambda^N \exp\left[-\lambda \sum x_i\right] \lambda^{\alpha-1} \exp[-\beta\lambda]$$

$$P(\lambda|x) \propto \lambda^{N+\alpha-1} \exp\left[-\lambda\left(\beta + \sum x_i\right)\right]$$

Some Applications

- Because this is a conjugate prior, we know that the posterior will follow a Gamma distribution.
- The kernel of the Gamma distribution has the form:

$$y^{\alpha-1}\exp[-\beta y]$$

- The kernel of our posterior has this form!

$$P(\lambda|x) \sim Gamma\left(N+\alpha, \sum x_i + \beta\right)$$

Some Applications

- Another less familiar one.
- Bayesian and maximum likelihood aren't all that different mechanically.
- Classic ML problems have close Bayesian analogues.
- N observations from uniform distribution $\sim Unif(0, \theta)$.

$$P(x_i|\theta) = \frac{1}{\theta}$$

- Uniform is not in exponential family, but still has a conjugate prior.
- Pareto distribution:

$$P(\theta) = \frac{k v_0^k}{\theta^{k+1}}$$

where

$$\theta \geq v_0 \ ; \ 0 \ o.w.$$

Some Applications

- Why does Pareto work here?
- θ must be greater than or equal to the sample maximum.
- MLE for θ is sample maximum, so there is similarity.

$$P(x|\theta) = \frac{1}{\theta^N}$$

$$P(\theta|x) \propto \frac{1}{\theta^N} \frac{kv_0^k}{\theta^{k+1}}$$

- Look at the kernel:

$$P(x|\theta) \propto \frac{1}{\theta^{N+k+1}}$$

- Conjugate, so posterior is Pareto:

$$P(\theta|x) = Pareto(k^* = N + k, v_0^* = max(x))$$

The Important One

- The previous applications actually occur in practice.
- The most common application is on the normal distribution.
- Recall that the normal distribution is a two-parameter member of the exponential family:

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2\right]$$

- With the normal, we can perform inference on μ and σ^2 separate or together.

Normal with Known Variance

- To start, let's look at a normal distribution with known variance.
- Looking for posterior of μ.
- μ should take values on full support of distribution.
- Put a normal prior on μ.

Normal Likelihood

- A key computation is the likelihood for a normal r.v.
- Start by expanding exponential:

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left[\frac{-1}{2\sigma^2}(x_i^2 - 2\mu x_i + \mu^2)\right]$$

- Now multiply:

$$P(x|\mu) = \left(\frac{-1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[\frac{-1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + N\mu^2\right)\right]$$

Prior on μ

- Define a normal prior on μ:

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right]$$

- Expanded:

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[\frac{-1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)\right]$$

- Match kernels:

$$P(x|\mu)P(\mu) \propto \exp\left[\frac{-1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + N\mu^2\right) + \frac{-1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)\right]$$

Completing the Square for Normals

- The normal is the conjugate prior for the mean. So, we know that the posterior for the mean will be normal.
- Consider the kernel of a general normal distribution:

$$\exp\left[-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right]=\exp\left[Ax^2+bx+c\right]$$

- We can solve for μ and σ^2 in terms of A and B:

$$-\frac{1}{2A}=\sigma^2\ ;\quad -\frac{b}{2A}=\mu$$

- If we can get the posterior kernel in this form, then we know the mean and variance of the distribution.

Posterior for μ

- Recall that the r.v. in our posterior is μ. So, we want to combine the kernels s.t.:

$$\exp[A\mu^2 + b\mu + c]$$

- We'll do this on the board because I don't feel like typing all this out.

$$P(\mu|x) \sim N(\mu^*, \sigma^{2*})$$

where

$$\sigma^{2*} = \left(\frac{N}{\sigma^2} - \frac{1}{\sigma_0^2}\right)^{-1}$$

and

$$\mu^* = \frac{\frac{\sum x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{(\sigma^{2*})^{-1}}$$

Posterior for σ^2

- Now let's assume that we know μ and we want to infer the value of σ^2.
- What is the conjugate prior for σ^2?
- Recall - what is the asymptotic distribution of variance?
- Variance is constrained to be greater than 0.

Posterior for σ^2

- Variance follows an inverse chi-square distribution.
- Chi-square is a special case of the gamma distribution.
- While inverse gamma is a distribution, it is often easier to work with precision (τ) - the inverse of variance.
- τ follows a gamma distribution.
- Gamma is conjugate to the precision of a normal distribution.

Posterior for τ

- Express the normal likelihood in terms of τ:

$$P(x|\tau) = \left(-\sqrt{\frac{\tau}{2\pi}}\right)^N \exp\left[-\frac{\tau}{2}\left(\sum x_i^2 - 2\mu\sum x_i + N\mu^2\right)\right]$$

$$P(x|\tau) = \left(-\sqrt{\frac{1}{2\pi}}\right)^N \tau^{\frac{N}{2}} \exp\left[\tau\frac{-\left(\sum x_i^2 - 2\mu\sum x_i + N\mu^2\right)}{2}\right]$$

- The prior on τ:

$$P(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}\exp[-\beta\tau]$$

Posterior for τ

- Match kernels:

$$P(\tau|x) = \tau^{\frac{N}{2}+\alpha-1} \exp\left[-\tau\left(\beta + \frac{\sum(x_i-\mu)^2}{2}\right)\right]$$

- The posterior for τ is:

$$P(\tau|\mu) = Gamma\left(\frac{N}{2} + \alpha, \beta + \frac{\sum(x_i-\mu)^2}{2}\right)$$

Bayes and MLE via Improper Priors

- The mean of the Gamma distribution is $\frac{\alpha}{\beta}$.
- Drive $\alpha \to 0$ and $\beta \to 0$ in the posterior of τ. What is the mean of the posterior distribution?
- MLE can be seen as equivalent to Bayes when we use a truly uniform prior and minimize MSE.
- However, this is a degenerate distinction for continuous distributions with infinite support.
- What is the uniform prior on $[0, \infty)$?
- $P(x) = 0 \; \forall \; x$ but $\int_x P(x)dx = 1$.
- Sometimes an improper prior doesn't matter, like the posterior for τ. However, it can matter if we want to do inference on the marginal likelihood or posterior predictive (more to come!).

Simultaneous Inference of τ and μ

- In general, we want to infer about both τ and μ.
- Conditional chain rule:

$$P(a, b|x) \propto P(x|a, b)P(a|b)P(b)$$

- If a and b are independent:

$$P(a, b|x) \propto P(x|a, b)P(a)P(b)$$

- Joint posterior:

$$P(\mu, \tau|x) \propto P(x|\mu, \tau)P(\mu|\tau)P(\tau)$$

Simultaneous Inference of τ and μ

- Simultaneous inference here requires an assumption that is a slightly weaker version of i.i.d. draws called *exchangeability*.
- In short, exchangeability essentially boils down to a thought experiment on labelling.
- Label N observations of x $x_1, x_2, x_3, ..., x_N$. Randomly permute these labels. If the results are the same, then the sample is exchangeable.
- When are things not exchangeable? Time series, panel data, etc.
- De Finetti's Theorem: exchangeable observations are conditionally independent given some latent variable to which an epistemic probability distribution would then be assigned.

Simultaneous Inference of τ and μ

- We already know the likelihood component of the normal model.
- We also know a prior on τ.
- Here, we need to define a prior on μ given τ, a normal with mean and precision:

$$P(\mu|\tau) = N(\mu_0, \kappa_0\tau)$$

- We'll come back to why this works in a minute.

Simultaneous Inference of τ and μ

- We use the same recipe as before:

$$P(\mu, \tau|x) \propto P(x|\mu, \tau)P(\mu|\tau)P(\tau)$$

- This math is kind of tedious and we should already have some intuition as to what the posterior distribution will look like. (Also, the result is kind of a letdown).
- What will be the form of $P(x|\mu|\tau)P(\mu|\tau)$?
- What happens when we multiply this by the prior on τ?

Simultaneous Inference of τ and μ

- The posterior has a normal-gamma distribution.
- This distribution is exactly what it sounds like - a normal distribution multiplied by a gamma distribution with parameters μ^*, κ^*, α^*, and β^*:

$$\mu^* = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_0 + N}$$

$$\kappa^* = \kappa_0 + n$$

$$\alpha^* = \alpha_0 + \frac{N}{2}$$

$$\beta^* = \beta_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \bar{x})^2 + \frac{\kappa_0 N(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}$$

Simultaneous Inference of τ and μ

- The math works out here, but takes a little longer than we have in this talk.
- This forms the *joint posterior*, $P(\mu, \tau|x)$.
- This is great, but we care more about the marginal posterior. What is the marginal density of the posterior of μ and τ?
- Integrate out the other parameter to get the marginal posterior.
- Once again, more math than we should do here.
- Any guesses about the marginal posterior forms?

Simultaneous Inference of τ and μ

$$P(\tau|x) \sim Gamma(\alpha^*, \beta^*)$$

$$P(\mu|x) \sim \mathcal{T}_{2\alpha^*}\left(\mu^*, \frac{\beta^*}{\alpha^*\kappa^*}\right)$$

Simultaneous Inference of τ and μ

- In the Bayesian paradigm, this is how the t-distribution appears - integrating out an unknown variance in a normal distribution.
- The variance is *compounded* into the uncertainty about the mean of the distribution.
- Let's go back to the prior on μ, $P(\mu|\tau)$. Here, we are making an appeal to the fact that the variance of the distribution of the mean is a function of the variance of the data generating process.
- In short, there's a CLT like argument being made here.

Predictive Distributions and Marginal Likelihoods

- Marginalizing is a key Bayesian computation.
- Marginalizing out parameters allows us to get distributions that we need for inference.
- What happens if we marginalize out all the parameters?

$$P(x) = \int_{\Theta} P(x|\theta)P(\theta)d\theta$$

- Does this look familiar?

Predictive Distributions and Marginal Likelihoods

- Assume that we have observed no data, but have the DGP and the prior forms.
- What is the probability of observing some value of x without any data?
- This is the prior predictive distribution.
- Think about what this implies - prior to collecting data, what is the probability distribution of $\widetilde{x}$?

Predictive Distributions and Marginal Likelihoods

- What about after we observe data, x?

$$P(\widetilde{x}|x) = \int_{\Theta} P(\widetilde{x}|\theta)P(\theta|x)d\theta$$

- This is called the posterior predictive distribution - after observing data, what is the probability distribution of $\widetilde{x}$?

Predictive Distributions and Marginal Likelihoods

- What about after we observe data, x?

$$P(\widetilde{x}|x) = \int_{\Theta} P(\widetilde{x}|\theta)P(\theta|x)d\theta$$

- This is called the posterior predictive distribution - after observing data, what is the probability distribution of $\widetilde{x}$?

Predictive Distributions and Marginal Likelihoods

- Why does any of this matter?
- We can answer questions about how well our model fits the data *a priori* and *a posteriori*.
- Using the posterior predictive distribution, we can examine how well the posterior distribution matches the observed distribution of the data - posterior predictive checks. This can also be used a modeling tool. Much, much more to come on this later.
- A more common question - conditional on our model choice, what is the probability of our data?

Predictive Distributions and Marginal Likelihoods

- Up until now, we've ignored the denominator of the Bayes machinery.
- This denominator is called the *marginal likelihood*. This tells us the probability that we would have observed our data given our prior beliefs.
- This is related to the *prior predictive distribution* in the following way for i.i.d. data:

$$\prod_{i=1}^{N} \int_{\Theta} P(x_i|\theta)P(\theta|\xi)d\theta$$

where ξ are the prior *hyperparameters*.

Predictive Distributions and Marginal Likelihoods

- On its own, this quantity (it's a constant, remember?) isn't too informative.
- However, when compared amongst models, it tells us which model is probabilistically more likely!
- Note that there are no assumptions about nesting, distribution of the ratio, etc.
- Consider two separate models M_1 and M_2 (on potentially different random variables), define the *Bayes factor* as:

$$BF_{1:2} = \log_{10}\left(\frac{\prod_{i=1}^{N}\int_{\Theta} P(x_i|\theta)P(\theta|M1)d\theta}{\prod_{i=1}^{N}\int_{\Theta} P(y_i|\theta)P(\theta|M_2)d\theta}\right)$$

Predictive Distributions and Marginal Likelihoods

- When the Bayes factor is negative, there is support for M_2.
- When the Bayes factor is positive, there is support for M_1.
- Differences in magnitude reflect differences in our belief that one model is better than the other!

Predictive Distributions and Marginal Likelihoods

- Estimating this quantity is often an incredibly difficult task.
- Approximate it or use numerical methods.
- More to come.

Multivariate Models

- Many distributions that we commonly think about have multivariate analogues.
- Not only is there variance to consider, but also *covariance* between observations.
- Let x_i be a random vector. How do we model this?
- A common approach is the multivariate normal distribution.

The Multivariate Normal Distribution

- The density:

$$P(x_i|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^k}}\frac{1}{\sqrt{|\Sigma|}}\exp\left[-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]$$

where k is the length of the random vector. μ is a $k \times 1$ vector of means and Σ is the associated $k \times k$ covariance matrix.

The Multivariate Normal Distribution

- Inference on the multivariate normal model is very similar to inference with a regular normal.
- As before, we're mostly interested in the case where both μ and Σ are unknown.
- Any guesses about the conjugate prior for μ?
- What about Σ^{-1}?

The Multivariate Normal Distribution

- As you probably guessed, the conjugate prior for μ is a multivariate normal.
- As you probably didn't guess, the conjugate prior for Σ^{-1} is a *Wishart distribution*.
- What is a Wishart distribution?
- Recall that variance is constrained to be greater than 0 and covariance has no constraints.
- Imagine a multivariate distribution that has gamma densities on the diagonal and normal densities everywhere else. That's a Wishart distribution.
- Hard to picture, but this is the conjugate prior to the covariance matrix in the multivariate normal problem.

The Multivariate Normal Distribution

- Just like in all of life, our time in this room is limited, so we're not going to go much further with this example.
- Unsurprisingly, the math is possible, but gets pretty tedious.
- Really, this model is just a bridge to the world of applied Bayesian statistics.

Bayesian Linear Regression

- Recall the linear regression model:

$$y_i \sim N(x_i\beta, \mathbf{V})$$

  where β is a p-vector of regression coefficients.
- How do we estimate β and $\mathbf{V}$ in a Bayesian way?

Bayesian Linear Regression

- Once again, the same machinery:

$$P(\beta, \mathbf{V}|x, y) \propto P(y|\beta, \mathbf{V}, x)P(\beta|\mathbf{V})P(\mathbf{V})$$

$$P(y|\beta, \sigma^2, x) \sim \prod_{i=1}^{N} \mathcal{N}_p(x_i\beta, \mathbf{V})$$

$$P(\beta|\mathbf{V}) \sim \mathcal{N}_p(\beta_0, \mathbf{V} \odot \mathbf{R}_0)$$

$$P(\mathbf{V}) \sim \mathcal{W}^{-1}(\nu_0, \iota_0)$$

Bayesian Linear Regression

- This is starting to get pretty difficult.
- What about if we have a binary dependent variable? Or a count? Or we believe that $\mathbf{V}$ is heteroskedastic?
- There are closed form solutions for some of these problems. But, the math gets too hard very quickly.
- Alas, all is not lost.
- Computers + Smart People = Solutions
- MCMC is the practical solution.